Advances in Cognitive Systems 1 (2012) 59–72 Submitted 2/2012; published 7/2012

Cognitive Foundations for a Computational Theory of Mindreading

Paul Bello PAUL.BELLO@NAVY.MIL
Human and Bioengineered Systems Division, Office of Naval Research, Arlington, VA 22203 USA

Abstract

Mindreading is the ability to understand both oneself and other agents in terms of beliefs, desires, intentions and other relevant mental states. This critically important ability has been implicated across a broad spectrum of human cognitive activities, including participation in dialogue, collaboration, competition, and moral judgment. This paper serves as a reflection on the kinds of strategies we can use to build a system capable of mindreading, given currently available resources in the relevant literature. After reviewing representative computational approaches on offer I will suggest a set of architectural mechanisms that could provide the flexibility required to build a robust mindreading capability for cognitive systems.

1. Introduction

As you were reading through the abstract, your mind engaged in a series of calculations designed to figure out what I meant in using the specific words and phrases that I used in writing it. Your capacity to "mindread" attempted to identify and ascribe to me a coherent set of beliefs, intentions, desires and other mental states that might have led me to write what I wrote. This powerful compulsion to predict and explain the behavior of other agents is not only active while reading text, but also while engaged in dialogue, while observing everyday human action, while competing or cooperating, when engaged with fiction of any type, and possibly when planning for the future or learning from past episodes.

The central cognitive activity involved in mindreading is the ascription of mental states from one agent to another. If Max observes Sally walking to the kitchen, he might infer that Sally is hungry, wants something to eat and will walk to the refrigerator because she thinks there is food inside. Max ascribes a number of mental states to Sally including her belief that food is in the fridge, that she desires to eat, and that she intends to walk to the fridge in order to get a snack. However, he likely does not ascribe other less relevant but logically possible mental states, such as Sally wanting to get something from the refrigerator and her believing that 89 is in the set of prime numbers. Although it seems odd to consider the latter as an example, such inferences are not only warranted but demanded on certain formal accounts of reasoning about beliefs.

The kitchen vignette described above is just one of a potentially infinite number of scenarios in which Max will be more or less successful at telling a respectable story about Sally's state of mind. How is this possible without Max having prior experience observing Sally when she is in a kitchen? Does Max maintain a huge collection of conditionals that roughly describes Sally's mental states? Does he assume that Sally will act rationally in the sense that her actions will be properly

conditioned on her presumed beliefs and desires? Can he reasonably think that Sally will behave much as he would in the same circumstances? Finally, if Max makes mistakes in attributing mental states to Sally, what kinds of mistakes will he make and under what circumstances will he make them? What sorts of mental representations and processes must Max be utilizing for all of this to happen so effortlessly? What kind of strategy should we pursue if we wish to build an intelligent system capable of the rich type of social cognition that Max displays?

This paper aims to lay groundwork for pursuing a research program in this important area, rather than offering a highly detailed solution. It is about the choices offered by the formalisms we have available at the moment, and whether they let us to tell a compelling story about how mindreading-capable agents might be built. The paper specifically focuses on representational and inferential requirements for the basic activities that comprise mindreading: *ascription* of mental states, *prediction* of behavior conditioned on these ascriptions, and *explanation* of observed behavior via *post hoc* ascription of mental states. Unfortunately, brevity requires us to focus on these core functions, rather than exploring the interesting connections between mindreading and other abilities, such as dialogue, moral judgment, and self regulation. After arguing that the tools we have for thinking about computational models of mindreading might not be up to the task, I present the very beginnings of what I believe to be firmer foundations for building such models.

2. Methodology and Modeling

In a recent *Cognitive Science* article, Cassimatis, Langley and Bello (2008) argued for three core criteria to be applied in the evaluation of models for higher-order cognition. These three criteria are *ability*, *breadth*, and *parsimony*. Generally speaking, by *ability* we meant the general capacity of a model to account for human-level competence with respect to the phenomena under investigation. By *breadth*, we meant that the model is capable of accounting for a variety (if not the preponderance) of phenomena-related results, including capturing competence-related trends across a sufficiently large space of human data. By *parsimony*, we meant that the model displays both ability and sufficient breadth without multiplying cognitive mechanisms (or representations) beyond the demands imposed by our most current data. As I shift discussion toward existing computational approaches to mindreading, I will argue that typically employed assumptions in both AI and computational cognitive science fail on at least one of these criteria.

As a matter of methodology, I am committed to not only giving a computational explanation of mindreading as a capacity, but also providing hypotheses for how it might be degraded or even fail outright. The strategy I adopt is to assume that error-prone mindreading is the result of cognitive systems that evolved for purposes other than mindreading and have since been re-purposed to the task of understanding other minds. One might argue that building a cognitive system that is prone to attribution errors seems wasteful or otherwise silly. I think that this remains to be seen. There are many types of social interaction where one agent benefits by having the ability to reason about the kinds of attribution errors made by another agent. For example, games like poker would be much less interesting for expert players if they were not able to apply a fairly rich model of errors to their advantage, even if they have no consciously accessible theory of attribution errors to draw from. The semantics of important social concepts like stereotyping would be difficult to capture in

a world without attribution errors. More generally, if we hope to have intelligent systems interact with human partners, it seems reasonable that the system be capable of predicting and responding to human errors in attribution. If my hunch is on target and mindreading is accomplished by means of multiple domain-general cognitive processes, we should expect that errors are the natural result of a cognitive system optimized for breadth. I argue later in the paper that the doorway to errors in mindreading is opened as a result of the parsimony of representation used to realize mental states in the cognitive system.

3. Computational Frameworks

I now turn attention to reviewing representative work in the computational literature on reasoning about the mental states of other agents. The frameworks I discuss are not and were never intended to be psychologically plausible or otherwise constrained by human data. I will argue that this is a fatal flaw. I will loosely refer to these as computational accounts of mindreading, but it might be more accurate to describe them as formalizations of epistemic reasoning, which is only one of the capabilities that we are interested in. I will attempt to analyze these approaches with respect to the ability, breadth and parsimony criteria discussed in the last section in order to facilitate comparison with my own approach to mindreading described later in the paper.

3.1 Logical Approaches to Epistemic Reasoning

Formal systems for representing and reasoning about the beliefs of agents are often expressed in the framework of epistemic logic (Hintikka, 1962). Knowledge and belief become modal operators that scope over formulae in the logical language. As an example of this approach, let us briefly describe the modal logic KD45, which is widely used in AI as a means of representing the knowledge and beliefs of agents. The letter **B** is the modal operator for belief, and relations of the form $\mathbf{B}\phi$ should be read as "Agent believes ϕ" where ϕ is some proposition.[1]

3.1.1 Ability, Breadth and Parsimony: An Analysis

Unfortunately, the purely logic-based approach is unsuitable for a psychologically plausible account of mindreading. The failure in this case is logical omniscience, or the general tendency for logics of this kind to require agents to over-generate inferences and to entertain irrelevant propositions, which I take as violating the ability criterion with respect to human-level mindreading competence. In general, KD45 does not take agents to be resource-bounded reasoners, which seems to be an obvious desideratum for building any sort of psychologically plausible computational artifact. KD45 contains the axiom *K*: $(\mathbf{B}\phi \wedge \mathbf{B}(\phi \rightarrow \psi)) \rightarrow \mathbf{B}\psi$, which compels an agent to believe the deductive closure of its beliefs. Furthermore, KD45 includes the axiom *N*: $\models \phi \rightarrow \mathbf{B}\phi$, which demands that all propositional validities are believed. Examples of such validities might be $\phi \vee \neg\phi$ or $\phi \rightarrow (\phi \vee \psi)$. Standard logical approaches do not differentiate between an agent's set of active beliefs and the

1. For our purposes, propositions are defined as truth-bearing descriptions of states of affairs. This contrasts with some other symbolic approaches to knowledge representation (e.g., production matching) in which "truth" is roughly equated with an element in working memory matching the antecedent of a production. Semantically speaking, explicit falsity is not a necessary feature of production systems.

slew of irrelevant tautologies that it is compelled to believe by way of *N*. For instance, Axiom *D*: $\mathbf{B}\phi \rightarrow \neg\mathbf{B}\neg\phi$ forbids agents from believing contradictions, which seems acceptable until it interacts with **K**. While it might be reasonable to assume that agents ought not to believe propositions of the form $\phi \wedge \neg\phi$, axiom **D** seems overly restrictive when ϕ is believed and $\neg\phi$ is a consequence of an incredibly long chain of reasoning resulting from the rule **K** over-generating consequences. Taken together, *D* and *K* impose an implausibly extreme commitment to consistency.

Axiom *4*: $\mathbf{B}\phi \rightarrow \mathbf{BB}\phi$ addresses the matter of positive introspection and agents believing whatever they believe. Axiom *5* is written as $\neg\mathbf{B}\phi \rightarrow \mathbf{B}\neg\mathbf{B}\phi$ and states that agents do not believe whatever it is they do not believe. Axiom *4* looks to be *prima facie* acceptable under certain conditions (Bello & Guarini, 2010), but *5* ought to be viewed with a healthy degree of suspicion, since it is often the case that a resource bounded agent will not know what it does not know. KD45's facilities for introspection also fall short of being able to represent and reason about ignorance as a first-class mental state. Limitations on representing ignorance make it difficult to utilize KD45 consistently in an intelligent system that needs to manage learning goals in a dynamic world.[2] There have been attempts to save these kinds of approaches from the unwelcome consequences of omniscience (Sim, 1997), but none have been widely adopted or implemented to date.

In general, logical approaches to mindreading fail to meet the parsimony criterion as well. The basic form of the argument is that it seems profligate to posit a different representation and special semantics for every propositional attitude we can conceive of. Should we consider a logic of ignorance, or of wishful thinking, or a logic of willfully violated obligations? Humans may well have something like logical theories that roughly describe the dynamics of such attitudes; however, I often wonder what kind of story to tell about how we arrive at them. The logical approach more or less assumes whatever semantic resources it requires as a convenience in order to explore conceptual terrain. To be admissible as the backbone of a theory of (human) cognition, such approaches must tell not only a compelling story about semantics, but also justify the kinds of representations they assume. One approach might be to provide a suitable reduction of complex attitudes (such as willful ignorance) into simpler attitudes like belief, intention, and desire. There have been various individual reductions of the sort that I describe, but little effort to systematize them or provide general principles for accomplishing them.

This brings our discussion to the breadth criterion. To illustrate my point, I consider Belief-Desire-Intention (BDI) models of agency (Rao & Georgeff, 1998) built within a broadly logical framework. The logical semantics associated with implementations of BDI theories require agents to be non-deceptive and helpful in ways that severely limit their applicability as off-the-shelf specifications for a socially competent system. Mental state attribution would be much less interesting if we were unable to ascribe willful ignorance, deceptive intentions, delusion, and other irrational attitudes. Far from being the unwanted byproducts of how our cognitive systems represent mental states, recognizing these attitudes in others serves as a soft constraint on future interactions. After all, most of us find little use in having extended discussion with the serially deluded, or in committing to a joint intention with an agent who has infelicitous intentions.

2. Although van der Hoek and Lomuscio (2004) attempt to address ignorance formally within the bounds of epistemic logic, as I read it their approach does not capture *total* ignorance, but rather captures being in the state $\neg\mathbf{B}\phi \wedge \neg\mathbf{B}\neg\phi$.

3.2 Probabilistic Approaches to Epistemic Reasoning

Aside from the classical logic-based approaches to belief ascription, a number of other implemented systems lay claim to being able to represent and reason about the beliefs of other agents. In recent years, approaches based largely on Bayesian and/or decision-theoretic commitments (e.g., Tenenbaum et al., 2011) have become fashionable due to their mathematical elegance and well-understood computational properties. Bayesian epistemology commits to the notion that rational agents have quantitative degrees of belief in statements, and that these degrees can be modeled using probability functions. As in most logical treatments of epistemic reasoning, probability theory entails that probability functions assign equal probability to logically equivalent statements and therefore equal degrees of belief. All tautologies generated by any sentential language P are believed with probability one, leading to the same kind of logical omniscience that characterized the logical language KD45 discussed earlier.

Some approaches to modeling non-omniscient Bayesian agents have been explored (Garber, 1983). For example, take a simple language with three variables, A, B, and C. One can adopt the strategy of treating C as the statement $A \rightarrow B$, such that $\Pr(B \mid A,C) < 1$. This seems to be a rather non-heroic way of rescuing Bayesian epistemology, since it demands that modelers need to generate *a priori* instances of every unknown relationship between elements of the language P, which is clearly undesirable. It is also a strange suggestion *vis a vis* cognitive architecture, since it suggests that all such statements are hidden somewhere in the subconscious, just waiting to be called up to be conditionally updated when the right kind of evidence presents itself.

Upon evaluation against the criteria of ability, breadth and parsimony, probabilistic approaches fare at least as poorly as classical logical approaches due to assumptions leading rational agents toward logical omniscience. Such approaches fail to meet the ability criterion as demonstrated by abundance of evidence for heuristics and biases in probabilistic judgments from the psychology of decision making (Tversky & Kahneman, 1974). In this sense, they may be good descriptions of mathematically sophisticated humans performing pen-and-paper exercises, but they are inadequate for capturing the judgments of untutored subjects. Since probabilistic approaches are fundamentally accounts of rational agency, they fail to meet the breadth requirement in exactly the same way as logical approaches. It is unclear how we might use these approaches to provide for the possibility of human error in decision making, the representation of total ignorance (e.g., a nonexistent or constantly changing domain of discourse), or the existence of attitudes that are essentially irrational. As for parsimony, probabilistic approaches fare nominally better than their logical counterparts, but that is where the good news ends for the Bayesian. The proponent of formal logics must explain why we ought to help ourselves to new semantics for mental states whenever we see fit, and without reduction to a core set of primitives. The Bayesian must explain why we do not need to help ourselves to virtually any semantics for propositional attitudes at all.

4. Cognitive Foundations for a Computational Model of Mindreading

Mindreading is particularly tricky business from an AI perspective, precisely because many of the standard assumptions employed in AI research – such as maximization of utility, rational belief updates, and the ability to compute the closure of one's beliefs seldom apply if we are to successfully

ascribe mental states to agents who we know are unconstrained by these principles. Below, I give a preliminary list of features to serve as a starting point for building *cognitive models* that are roughly consistent with the data available on human-level mindreading abilities. The list is structured to pull apart representational considerations from those about cognitive processing and generally desirable mindreading-related abilities.

- Representational requirements
  - Perspectives/simulations as a deep representational commitment
  - Ability to consider nested mental states
  - Ability to represent and reason about ignorance
- Processing requirements
  - Lazy inference and/or incrementality
  - Flexible methods for populating different perspectives/simulations
  - Mutable ascriptions from mindreader to target.
- Desirable features
  - Ability to cope with mispredictions and incorrect ascriptions
  - Consistency with data on human development
  - Explanation of the relationship between mindreading and introspection

As I present my own work on mindreading in the remainder of this section, I will reference many of the items listed above.

4.1 Cognitive Architectures, Modularity, and Mindreading

One of the problems afflicting the formal approaches covered in the previous section is what I will call *homogeneity*. By this I mean that their proposed representations and processing mechanisms lack the flexibility needed to account for the complexity suggested by recent empirical findings. I adopt a strategy of *constrained heterogeneity* that implements mindreading as a principled collection of interacting modules.[3] I contend that the constrained heterogeneity required by a full account of mindreading can be found in theories of the human cognitive architecture. Such theories make claims about the representations, processes, and integration mechanisms that underlie human mental life (Langley et al., 2009). In practice, cognitive architectures are often implemented as a collection of interacting modules which implement many of the distinctions we find in the psychological literature. For example, some architectures implement distinctions among long-term and working memory (Langley & Choi, 2006), between action-related and non-action-related cognition (Sun & Zhang, 2003), and between mechanisms for dealing with different types of cognitive content (Cassimatis, 2006). The modular nature of cognitive architectures let us tie some of the variance we find in the empirical data to mechanisms defined at a finer grain size. Finally, modularity lets us constrain our solution in such a way that it is generally consistent with what mechanisms we know developing children to possess, rather than providing a homogeneous competence-only account that bears little relation to plausible cognitive capabilities or related performance.

3. By this, I do not mean the classical Fodorian modules that are sometimes invoked in service of explaining mindreading abilities (Leslie et al., 2004). I simply mean a collection of representational and processing elements defined at a finer grain than those typically used by the formal approaches outlined in Section 3.

As noted earlier, the three main capabilities to be accounted for in mindreading are *ascription* of mental states, *prediction* of behavior based on ascriptions, and *explanation* of observed actions by way of *post hoc* ascriptions. Even though I have used a specific cognitive architecture in my explorations in modeling mindreading, nothing that follows precludes the use of different architectures or appropriately constrained integrated intelligent systems. At an abstract level, architectures for mindreading will likely require some basic capabilities for considering alternate states of affairs, for reasoning about identity, for forward inference, and for subgoaling. These basics follow directly from the task requirements imposed by ascription, prediction and explanation, respectively.

If we want to adopt the commonplace ascription-via-simulation strategy, in which A reasons about B's beliefs by imagining himself-as-B, there must be some representational provision for considering alternate states of affairs and reasoning about identity (i.e. himself-as-B). Beliefs and other mental states are *opaque* in the sense that they are essentially private. Logically speaking, if Clemens = Twain, Agent A can believe x is Clemens without believing x is Twain. Agent B can believe that x is Twain without believing x is Clemens, and so on. Keeping A's beliefs separate from B's beliefs requires a deep commitment to agents having different perspectives on the world. The requirement for forward inference seems to be obvious in light of the need to predict behavior. Once ascription is complete, we must match against available plans in order to generate a set of predictions about possible behaviors. Finally, explanation would seem to sometimes involve observing or inferring an action, and then creating subgoals to identify potential causes.

4.2 A General Framework

My collaborators and I have offered up a theory of mindreading grounded in the domain-general operations of a computational cognitive architecture (Bello et al., 2007). In past work, we have shown how both aspects of introspection and third person ascription are reducible to a substrate of domain-general representational primitives and processing elements that include mental simulation of counterfactual worlds, reasoning about identity, reasoning about categories, and applying conditional rules. While this sounds like quite a lot of mechanism, all of these abilities seem to be in place in typical two-year olds, and none of them implies any commitment to innate modules or special representations for mindreading. We take mental simulation to be a critical operation for the ascription of beliefs, which according to our theory proceeds in six steps:

Categorize: Categorize the other entity as an agent

Instantiate: Construct a counterfactual world C where self = other is true;

Populate: Select a relevant subset Φ of the self's candidate beliefs to use in populating C;

Discriminate: Detect differences between the self and the other with respect to Φ;

Amend: For each difference detected, override the truth values of self-related propositions in favor of other-related propositions; and

Simulate: Proceed with inference in C and predict the other's behavior.

As a basic notation, I will use expressions like *Relation*(e_1, . . ., *worldname*) to express propositions that consist of a relation over a series of arguments. The final argument denotes the "world" in which the relation takes a truth value in the range true, uncertain, false ({T, U, F }).

4.3 Inheritance, Overrides and Mindreading

When we mentally simulate an alternate world, we would like as many facts as possible about the real world to stay the same. If I assume in simulation that the apple is at loc1 in w1, I would also like to have the fact that the apple is red available to me in w1. Let us call the real world "R." I know the apple is red in R. What we need is some way to connect R to w1 so that information from R becomes available for use in w1. This *inheritance* process is crucial in explicating our particular account of mindreading. Formally, we say that two worlds are relevant to one another when the basis of one world is fully contained in the basis of the other.[4] The relevance relation is transitive, making it possible to reason about nested beliefs using our framework.

Relevance: (Basis(?w1) $\subseteq$ Basis(?w2)) $\rightarrow$ RelevantTo(?w1, ?w2)

Transitivity: RelevantTo(?w1, ?w2) $\wedge$ RelevantTo(?w2, ?w3) $\rightarrow$ RelevantTo(?w1, ?w3)

The basic form of a rule that enables inheritance can now be written:

I_{basic}**:** ?Relation(?e$_1$, . . ., ?w1) $\wedge$ RelevantTo(?w1, ?w2) $\rightarrow$?Relation(?e$_1$, . . ., ?w2)

This inheritance rule allows for *hypothetical* reasoning. With it, we can consider hypothetical worlds in which nothing that we assume contradicts anything we know about the actual world. Applying I_{basic} lets us migrate things that we know about the real world into the hypothetical world under consideration.

However, simulation-based theories of mindreading rely centrally on the notion of entertaining counterfactuals rather than hypotheticals. Counterfactual worlds are predicated on propositions we know to be false in the real world. In our framework, counterfactual worlds are no different from hypothetical worlds or other sorts of simulated worlds in terms of underlying mechanism. The difference lies in the definition of their relationships to their parent worlds via inheritance. Entertaining a counterfactual world w2 requires a basis proposition ?CfAtom(?e$_1$ . . ., ?w1) in w2 such that RelevantTo(?w1, ?w2) and $\neg$?CfAtom(?e$_1$ . . ., ?w1). For example, if I know it is sunny in the real world, but I would like to consider a counterfactual world cf where it is not sunny, I would have: Basis(R) = { }, Sunny(E, R), and Basis(cf) = {$\neg$Sunny(E, R)}.

Notice that the proposition in the basis of cf has a world argument that references R (or the parent world, more generally). This captures the notion that the counterfactual world is *about* a proposition in the parent world. If we were to employ the inheritance rule I_{basic} to Sunny(E, R), we end up with an immediate contradiction in cf. To avoid this, we write a new inheritance rule as a *soft* constraint of the form:

$I^{\uparrow}_{cf}$**:** ?Relation(?e$_1$, . . ., ?w1) $\wedge$ RelevantTo(?w1, ?w2) $\wedge$ IsCounterfactualWorld(?w2, ?w1) $\rightarrow_{cost}$?Relation(?e$_1$, . . ., ?w2),

4. The basis of a world is the set of assumptions on which that world is based. Scally et al. (2011) provide details on reasoning about second-order beliefs using the definition of relevance provided in this paper.

where *cost* takes a value in the range (0,1), but typically in the very upper end of the range (e.g., 0.95).[5] Any proposition that is the consequent in a soft constraint of this form will not have a truth value in {T, F}, but will instead be considered uncertain and subject to further inference. The uncertainty about these propositions can then be further resolved by backtracking search. In our framework, backtracking involves simulating hypothetical worlds in which the uncertain proposition is true and others in which it is false till a truth value is settled on.

4.3.1 Downward Inheritance

So far, I have explored cases where propositions in a parent world are available to their children, which we call *upward* inheritance. In contrast, *downward* inheritance involves inheriting propositions from a child world downward into the parent world:

$I^{\downarrow}_{cf}$**:** ?Relation(?e$_1$, . . ., ?w2) $\wedge$ RelevantTo(?w1, ?w2) $\rightarrow_{cost}$?Relation(?e$_1$, . . ., ?w1).

In similar fashion to the example of counterfactual reasoning given above, counterfactual conclusions generated in a child world can migrate downward to the parent world as uncertain, with a backtracking search process resolving the uncertainty where possible.

Most of the time, there will be propositions in the parent world that suppress the counterfactual consequences inherited downward, however on some occasions this might not occur. Downward inheritance has the relatively unsettling implication that, in the absence of having real-world information to the contrary, content generated within counterfactual simulations becomes available to the simulating agent's set of beliefs about the real world. Now, this sounds a bit far-fetched until we think about emotional engagement with fiction or children engaged in pretense. Information generated in these worlds must be made available to action selection and generation. Similarly, when we plan using hypothetical worlds, or when we generate plans prior to ever encountering any real-world stimuli, we must have a way of inducing conditional actions. The downward inheritance mechanism is one such method for achieving this sort of functionality.

Downward inheritance also provides a way to deal with deception in a way that BDI-style frameworks fail to do. If I have information in the real world that my interlocutor often lies to people, and then he tries to inform me of the proposition *P*, one of the inferences I would normally make is that he believes that *P*. In our framework, this would amount to *P* being true in the counterfactual world where I am him, and downward inheritance makes *P* immediately available to my own set of beliefs. But since *P* inherits back into my own beliefs as uncertain, and the knowledge I have about my interlocutor leads me to believe he is lying (and actually $\neg P$), I am in a situation where my convictions about $\neg P$ trump the information made available through downward inheritance. Interestingly enough, our commitment to these rather counterintuitive mechanisms has recently been vindicated. In a recent *Science* article, it has been found that entertaining the beliefs of other agents (even if they are false) influences reaction time measures for simple tasks in a way consistent with our downward inheritance hypothesis (Kovacs et al., 2010). Less than a month after publication, I was able to build a model using our framework that is consistent with their findings without altering our existing theory of mindreading (Bello, 2011).

5. One approach to using costs in processing soft constraints is presented by Scally et al. (2011).

4.3.2 Overriding Default Ascriptions

The account I have presented relies centrally on the notion of *default ascription*. This consists of assigning one's own mental states to a target agent when no specifics about the target's mental life are known *a priori*. But what happens when I as the mindreader know that you as the target have a false belief? When I simulate the counterfactual world in which I am you, I have to suppress egocentric attributions (e.g., my own beliefs about the world) and promote attributions that I know to be consistent with your beliefs. Two additional rules make self-indexed beliefs look like other-indexed beliefs in the context of the counterfactual world where self = other:

O$^+$: IsA(?other, Agent, E, ?w1) $\wedge$ IsA(self, Agent, E, ?w1) $\wedge$?Relation(?other, . . ., ?w1) $\wedge$ $\neg$?Relation(self, . . ., ?w1) $\wedge$ RelevantTo(?w1, ?w2) $\wedge$ Same(self, ?other, E, ?w2) $\rightarrow_{cost}$?Relation(self, . . ., ?w2)

O$^-$: IsA(?other, Agent, E, ?w1) $\wedge$ IsA(self, Agent, E, ?w1) $\wedge$ $\neg$?Relation(?other, . . ., ?w1) $\wedge$?Relation(self, . . ., ?w1) $\wedge$ RelevantTo(?w1, ?w2) $\wedge$ Same(self, ?other, E, ?w2) $\rightarrow_{cost}$ $\neg$?Relation(self, . . ., ?w2)

These rules are quite general, and cover any differences between the self and the other that might be salient. They predict that costs will accrue linearly when mindreading increasingly dissimilar targets. This prediction has recently been lent some support by results reported in (Tamir & Mitchell, in press). Learning additional overriding constraints associated with individual targets is possible within the framework I have described, but brevity precludes a detailed discussion.

4.3.3 Controlling Inheritance

Implementing inheritance as soft constraints offers critical flexibility since it is possible for the numerical costs associated with each constraint to be externally influenced. One way this might happen is via the influence of other events occurring in the architecture during an episode of mindreading. For example, multitasking during episodes of mindreading has been shown to increase egocentric misattributions of mental states from a mindreader to a target agent. Let us assume that costs on inheritance constraints reflect the amount of attention we pay to self/other differences. If these costs are recomputed periodically with respect to other events occurring in the architecture, we could imagine that multitasking would lower sensitivity to differences and lead to egocentric misattributions.

Another way that costs on constraints can be influenced externally is via explicit judgments made by the mindreader. It has been shown that the degree to which mindreaders self-identify with target agents (usually measured by questionnaires) affect the number and quality of attributions made. This should not come as a surprise to any of us. We typically have richer models of the mental lives of close others than we do of perfect strangers. Recent studies indicate that we pay less attention to differences between ourselves and those identified as close others than we do when considering dissimilar others (Savitsky et al., 2011). If quantified, such self/other similarity measures could modulate costs on inheritance constraints that would reproduce such patterns.

Finally, it should be clear that giving a complete account of inheritance involves solving some form of the relevance problem in AI. My account of mindreading predicts that whatever cognitive

mechanisms enable relevance calculations for the mindreader should be operative in the selection of inheritance rules to consider during mindreading.

4.4 Ability, Breadth and Parsimony

The account I have presented above fares substantially better on the ability, breadth and parsimony metrics than those we reviewed earlier. Fully worked examples of mindreading using the assumptions in this section have been provided elsewhere, especially in Bello et al. (2007), Bello (2011), and Scally et al. (2011). In the present account, an agent A representing the belief *BName* of an agent B consists of: Basis(Bworld) = {Same(A, B, R)} and *BName*(e_1, . . ., Bworld). This conception of belief does not demand any special representations or domain-specific processes, but relies on the domain-general ability to simulate counterfactual worlds. Performance on counterfactual reasoning tasks has been consistently seen to be correlated to performance on mindreading tasks (Riggs & Peterson, 2000). Coincidentally, recent data on the ability of children to reason about false beliefs places the emergence of such reasoning in roughly the same developmental timeframe as when children start to spontaneously engage in pretense. Both activities are served by the same cognitive mechanisms including the simulation of worlds and controlling the flow of content between worlds via inheritance relationships. The downward inheritance of content from child worlds into their parents is an assumption we make about counterfactual reasoning that explains some of the puzzling phenomena we see in the empirical data. Arousal generated by engagement with fiction and altercentric ascription errors are side effects of assumptions we make about counterfactual reasoning, rather than being hand-coded features.

Lazy inference coupled with the aforementioned mechanisms provide a highly parsimonious account of how even pre-verbal infants can entertain the mental states of others. The considerable flexibility of inheritance rules provide for the possibility of both egocentric and altercentric attribution errors, such as those documented by Kovacs et al. (2010) and modeled by Bello (2011). Reasoning about deceptive agents and agents with false beliefs is possible through the process of overriding uncertain propositions that move from one world into the other, depending on the respective direction of inheritance. The assumptions I have made to account for basic mindreading capabilities are widely thought to be in the possession of infants between 12 and 18 months of age. Most of what I have assumed are mechanisms crucial for reasoning about the physical world, such as basic capacities for spatial, temporal, categorical, and causal inference. New developmental studies (e.g., Onishi & Baillargeon, 2005) that employ sensitive non-verbal measures of false belief understanding are finding competence with mindreading during roughly the same time frame as when infants begin to spontaneously entertain pretenses, lending support to our assertion that belief and pretense overlap to some degree at the implementation level.

5. Conclusions

Mindreading represents one of the most complicated and interesting cognitive activities in which we routinely engage. As such, we ought to take it seriously as a major desideratum in the development of cognitive systems. The overall aim of this paper has been to illustrate the complexities of mindreading and the relative difficulty in trying to account for them using assumptions that typify

standard techniques in AI. Many of these assumptions are prescriptive by their nature, and enforce constraints on rationality that are rarely satisfied during real-world episodes of mindreading or even during controlled studies performed in laboratory settings. I have argued that a deflationary account of the mental states of others consisting primarily of counterfactual simulations and inheritance explains the close relationship between performance on mindreading tasks and data on entertaining pretenses.

Standard accounts of propositional attitudes assume a sharp delineation between mental states, usually related to the kinds of actions that they tend to motivate. At best I think we have seen that this assumption is questionable, and at at worst it seems wrong. When taken to unreasonable extremes, it seems as if totally decoupling mental states from one another at the level of implementation makes it difficult to explain engagement with fiction, empathy, wishful thinking, self-deception, pretense, delusions or hallucinations. While some theorists see these as unfortunate outliers, I have argued that mindreading-enabled systems should be able to recognize them in others and modify their interaction strategies accordingly. Having a system that initially is capable of exhibiting all of these behaviors and using simulation to recognize them in others seems to be a reasonable alternative to the rather ugly option of trying to axiomatize them in service of reasoning about them.

I have further argued that inheritance rules implemented as soft constraints lets us fit a wide swath of data on mindreading than spans the gap between totally incorrect and perfectly correct attributions. Under assumptions of unlimited inferential resources, this range of attributions accounts for systematic mispredictions and perfectly rational epistemic inference alike. There is much work to be done to flesh out my suggestions into a robust implementation. While the representation of inheritance as soft constraints allows for variance in the attribution process, it is unclear how to systematically link costs on constraints to other features of ongoing cognition, including explicit judgments and resource limitations in the cognitive system. I have also intentionally left the discussion of learning new inheritance constraints from successful and unsuccessful episodes of mindreading as an open issue. The issue of whether or not such learning is automatic or intentionally initiated remains open, and computational expressions of the learning process are equally undeveloped. The influence of affect, emotions, feelings, and physiological variables on inheritance is completely unexplored in this paper, as is the question of how to reason when uncertain about the mental states of the target or when knowing the target to be uncertain about a proposition of interest. I have also not spent any time on the relationship between third person mindreading and introspection. In short, this paper has barely scratched the surface, but I hope the suggestions that I have provided will serve as a good starting point for researchers who are interested in accounting for both mindreading competence and architecture-level performance in a parsimonious way.

Acknowledgements

I am indebted to Pat Langley for his careful reviewing of earlier versions of this document and his subsequent suggestions, which have dramatically improved the structure of the paper. I would also like to thank Will Bridewell, Selmer Bringsjord and Bertram Malle for their trenchant commentary on the conceptual contents. Finally, my deepest gratitude goes out to Nick Cassimatis for his continued mentorship and seminal contributions in the development of these ideas.

References

Bello, P. (2011). Shared representations of belief and their effects on action selection: A preliminary computational cognitive model. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2997–3002). Boston, MA.

Bello, P., Bignoli, P., & Cassimatis, N. (2007). Attention and association explain the emergence of reasoning about false belief in young children. *Proceedings of the Eighth International Conference on Cognitive Modeling* (pp. 169–174). University of Michigan, Ann Arbor, MI.

Bello, P., & Guarini, M. (2010). Introspection and mindreading as mental simulation. *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society* (pp. 2022–2028). Portland, OR.

Cassimatis, N. (2006). A cognitive substrate for human-level intelligence. *AI Magazine*, *27*, 45–56.

Cassimatis, N., Bello, P., & Langley, P. (2008). Ability, parsimony and breadth in models of higher-order cognition. *Cognitive Science*, *33*, 1304–1322.

Garber, D. (1983). Old evidence and logical omniscience in Bayesian confirmation theory. In J. Earman (Ed.), *Testing scientific theories*. Minneapolis, MN: University of Minnesota Press.

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

Kovacs, A., Teglas, E., & Endress, A. (2010). The social sense: Susceptibly to others' beliefs in human infants and adults. *Science*, *330*, 1830–1834.

Langley, P., & Choi, D. (2006). A unified cognitive architecture for physical agents. *Proceedings of the Twenty-First National Conference on Artificial Intelligence*. Boston, MA: AAAI Press.

Langley, P., Laird, J., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.

Leslie, A., Friedman, O., & German, T. (2004). Core mechanisms in theory of mind. *Trends in Cognitive Science*, *8*, 528–533.

Onishi, K., & Baillargeon, R. (2005). Do 15-month old infants understand false beliefs? *Science*, *308*, 255–258.

Rao, A., & Georgeff, M. (1998). Decision procedures for BDI logics. *Journal of Logic and Computation*, *8*, 293–343.

Riggs, K., & Peterson, D. (2000). Counterfactual thinking in pre-school children: mental state and causal inferences. In P. Mitchell and K. Riggs (Eds.), *Children's reasoning and the mind* (pp. 87–99). Hove, UK: Psychology Press.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Sawnson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology*, *47*, 269–273.

Scally, J., Cassimatis, N., & Uchida, H. (2011). Worlds as a unifying element of knowledge representation. *Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems* (pp. 280–287). Arlington, VA: AAAI Press.

Sim, K. (1997). Epistemic logic and logical omniscience: A survey. *International Journal of Intelligent Systems*, *12*, 57–81.

Sun, R., & Zhang, X. (2003). Accessibility versus action-centeredness in the representation of cognitive skills. *Proceedings of the Fifth International Conference on Cognitive Modeling*. Bamberg, Germany.

Tamir, D., & Mitchell, J. (in press). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General*.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Structure, statistics, and abstraction. *Science*, *331*, 1279–1285.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

van der Hoek, W., & Lomuscio, A. (2004). A logic of ignorance. *Electronic Notes in Theoretical Computer Science*, *85*, 1–17.