Folk Concepts and Cognitive Architecture: How We Believe

Paul F. Bello
Office of Naval Research, Arlington, VA USA 22203
paul.bello@navy.mil

Abstract

It seems painfully obvious that whatever constitutes the adult concept of believing has something to do with our experience as believers. That is to say that our *folk concept* of belief supervenes on the cognitive and affective processes that undergird our ability to both believe and to ascribe beliefs to others. The nature of this relationship between concepts and the cognitive processes to which they refer is very much an open question. In any case, there has been little to no work on computational cognitive models of belief and belief ascription allowing for the systematic exploration of this relationship. This paper begins to sketch out the very beginnings of a cognitively-inspired process model for belief ascription that is broadly consistent with results spanning the cognitive sciences.

1 Introduction

Folk concepts play an integral role in our lives. Without them, it's unclear that many of the central day-to-day activities we involve ourselves in would even be possible. This is especially true of social interactions, where we often predict and explain behavior by invoking mental-state terms. We often find ourselves saying things like "Johnny was hungry and went downstairs to the fridge because *he thought there was leftover pizza*; but once he opened the door *he realized Jane must have eaten it* while he was out at the store." By a *folk concept* of *x*, I mean the collection of distinctions we are apt to make regarding *x* and the role that *x* plays in thinking and doing. For example, we typically distinguish intentional from unintentional (or accidental) behavior, usually be invoking agentive causes, some notion of premeditation and/or desire for said agent to bring the behavior about. Since it has been the most systematically studied folk concept, this paper focuses squarely on belief. But what does it mean to believe, anyway? How does it differ from accepting, or knowing, or pretending about some proposition of interest? The traditional answer to this age-old question typically invokes some notion of *role-functionalism*. Beliefs are what they are by virtue of the unique part that they play in generating and/or explaining downstream mental or physical behavior. This is functionalism about the mind, *par excellence*. But this relatively simple story has something of a complicated relationship with the data and often with our own intuitions.

For a motivating example, let us briefly consider the relationship between imagining or pretending and believing, which has been written about by a variety of scholars (Langland-Hassan in press; Nichols 2004; Gendler 2007). One phenomenon that muddies the water for the role-functionalist concerns the real effects of fictional immersion on normal humans. We cringe during horror movies, and we convince our children that sleeping in their room is safe after a particularly engaging episode of battling imaginary monsters under their beds. How can this be? The role-functionalist would claim that beliefs have a causal relationship with real-world action that imaginings or pretending

somehow lack. It is by virtue of these causal roles that we bin certain kinds of mental representations as instances of believing and others as instances of imagining. But the data suggests otherwise – it is by virtue of imagining that we sometimes act as we do, even if said actions do not fit cleanly into the individuation scheme imposed by certain kinds of functional role semantics. Moreover, in some cases like seeing horror movies we are certainly apt to make statements about how these kinds of fictions make us behave, locating real-world action somewhere in our folk-concept of fiction, which of course, doesn't seem to distinguish between belief and fantasy in a meaningful way.

What all of this suggests to me is that functional roles, at least when strictly expressed at the level of mental-state terms, are not the *sine qua non* for uniquely specifying the input/output mapping for mentality. The examples I referenced earlier regarding the relationship between believing and imagining speaks to some kind of interaction between the two at the level of cognitive processes. The purpose of this paper isn't to explore the pretense-belief connection in substantial detail, but rather to suggest that our folk concepts might be better explored in terms of cognitive process models, since they might provide explanatory resources that coarser conceptual definitions cannot. .

2 Toward a Cognitive Model of Belief Ascription

Defending and fully elaborating a cognitive model of belief ascription within the scope of such a short paper would be impossible, so I do not attempt to do so here. Instead, I will present a high-level architectural sketch of what such a model might look like, motivating each of its components in a relatively cursory fashion in the interest of brevity. But before I do so, it's worth diagramming the model so that it can be referred to as the discussion continues to unfold. The general picture I have in mind is shown in figure 1:

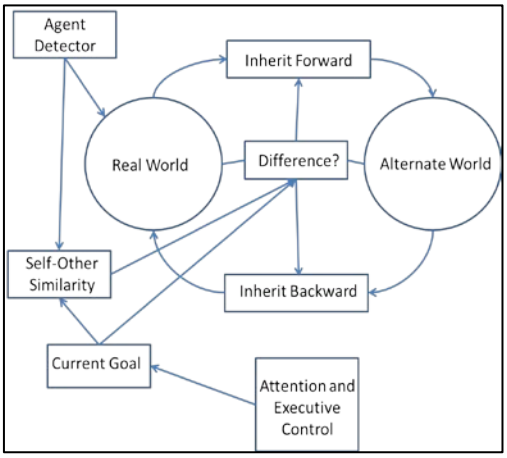

Figure 1: A Process Model of Belief Ascription

The model represented in the figure above requires some substantial unpacking to make sense. In prior work, my colleagues and I have adopted a largely *simulation-theoretic* take on belief ascription (Bello et al 2007; Bello & Guarini 2010; Bello 2011). Simulation theory roughly consists in the view that when I want to think about what you might be thinking, I adopt your perspective on the world, and use my own set of beliefs and practical reasoning mechanisms as an approximation of the contents of your mind. To be successful at predicting and explaining behavior, this strategy depends on the simulating agent to have a congruent perspective to that of the agent being simulated, at least in the context of their interaction. Naturally, the question of what populates these perspectives for a

given interaction is an utterly open one, and takes us much too far afield to consider in any depth whatsoever. Needless to say, if for a given interaction context *C*, the simulator and simulated agents have substantively different perspectives, the former needs to suitably suppress his own view on the world and adopt the perspective of the simulated agent in order for prediction or explanation to be successful.

2.1 Motivating the Model

The basic algorithm that the model in figure 1 implements runs something like the following:

1. **Categorize**: Automatically categorize the other entity as an agent and monitor line-of-sight or observe an action of interest taken by the agent in the real world.
2. **Instantiate**: Construct an alternate world *w* representing the perspective of the other;
3. **Discriminate**: Detect differences between the self and the other with respect to current interaction goal (if any exist);
4. **Populate**: Select a relevant subset of the self's candidate beliefs to use in populating *w*;
5. **Amend**: For each difference detected, override the truth values of self-related propositions in favor of other-related propositions
6. **Infer**: Proceed with inference in *w* and predict or explain the other's behavior.
7. **Supress**: Send inferred information in *w* back to the real world. If in conflict with anything currently known, amend and/or suppress.

To begin, the model in figure 1 rests on the notion that an agent *A* has a set of propositional content associated with its beliefs about the real world. Let us assume that these contents are located in the circle labeled **Real World**. Since the ascription of belief by way of mental simulation requires the construction of other worlds representing the contents of other agents' minds, we have a corresponding circle called **Alternate World** and connections between them that I address momentarily. In brief, the **Real World** represents *A*'s perspective on the world, and **Alternate World** captures *A*'s take on the perspective of some other agent *B*. The evidence for the automatic construction of perspectives in the presence of other agents comes by way of multiple studies of which (Kovács et al. 2010; Surtees & Apperly 2012) are two examples. Both of the aforementioned results suggest that the mere presence of agents in a task environment drive the spontaneous adoption of perspectives, even in the absence of explicit instructions to adopt them. The model in figure 1 accounts for steps 1 and 2 by way of the **Agent Detector** box, and its connection to the **Alternate World** shown on the right.

Evidence for steps 3 and 4 come to us indirectly through studies of so-called *egocentric attribution errors* (Birch & Bloom 2007; Keysar et al. 2003). An egocentric attribution error involves agent *A* mistakenly attributing its own beliefs to another agent *B* without proper warrant. Keysar et al. (2003) presented a series of objects in a grid, some of which shared properties such as color or shape, but differed on other dimensions. Behind the grid stood a director, who could only see a subset of the objects in the grid. The director gave commands to subjects of the form "move the <property-description> <object-name> <up/down/left/right>" (e.g. "Move the red ball to the left"). Manipulations were introduced such that the discrepancy in perspective between the subject and the director led the subjects to egocentrically over-ascribe beliefs to the director. In one example, the subject sees three balls of different sizes in the grid. Some items in the grid are visible only to the subject and not to the director. In this case, the smallest of the three balls is not able to be seen by the director. During a trial, the director says to the subject "move the small ball up." In order for the subject to successfully do so, he must take account of the fact that the small ball seen by the director is actually the medium-sized ball as he sees it. Cases like these typically generate egocentric errors, or

at minimum generate extended reaction times. Subjects have a hard time suppressing their own beliefs about which object is the smallest without devoting effort to the process.

This kind of study substantiates step 4, where some subset of agent *A*'s beliefs populate agent *A's* model of agent *B's* mind. In a related study, Savitsky and colleagues used the same task, but manipulated the interpersonal similarity relation between the subject and the director (Savitsky et al. 2011). Subjects participated in pairs, and were asked to rate on a quantitative scale how close they felt to their counterpart from "complete stranger" to "best friend." Higher ratings of interpersonal similarity resulted in more frequent commission of egocentric errors. The conclusion to be drawn is that ascription depends crucially on how you construe your social interlocutor. If construed as being highly similar, little attention is allocated to monitoring differences between self and other, and the likelihood of egocentrism increases. On the other hand, being in the presence of a complete stranger directs attention to differences, and results in more careful perspectival bookkeeping, which we associate with step 4 in the algorithm given above. Steps 3 and 4 are associated with the operation of the **Self-Other Similarity**, **Current Goal**, **Attention & Executive Control**, **Difference** and **Inherit Forward** functions in figure 1.

Step 5 describes the process by which perspectives are rectified. Once it has been established that there are substantive differences between what the ascriber believes and what the target of ascription believes, a suppression process involving the effortful direction of attention works to arrange the **Alternate World** in a manner consistent with the detected differences. Recently produced evidence for *egocentric anchoring and adjustment* suggests that when making ascriptions to similar others, ascribers seem to start by assuming commonality between their own perspectives and that of the target, adjusting away as differences are accounted for (Tamir & Mitchell in press; Lin et al. 2010). In the study by Tamir and Mitchell, subjects were asked to report their own attitudes on a topic, and those of a similar or dissimilar other. Subjects were selected by virtue of their political leanings, and asked to self-assess on a series of their likes, dislikes and other attitudes regarding a variety of topics. Descriptions of two prototypical others were provided to the subject in short paragraphs – one liberal and one conservative, after which subjects judged how strongly the targets would endorse statements about the same topics they previously self-assessed on. Both the self-assessment and other-endorsement scales were aligned with one another to provide a fine-grained measure of self-other discrepancy. Reaction times for judging the attitude of similar others linearly trended with the self-other discrepancy measure. This is to say subjects seem to have used their own attitudes as anchors and adjusted away with respect to the description of the similar other. In the case of dissimilar others, reaction times were (1) larger, and (2) unrelated to self-assessments.

I shy away from making comments about step 6, because there seems to be no clear consensus regarding just how our practical reasoning system does what it does, and what it means to ascribe *knowledge-how* to target agents in order to describe idiosyncratic inference rules that they might be using. Even if we ascribe *knowledge-how* of this form to other agents, we have only the haziest idea about how such ascriptions figure into the downstream prediction or explanation of the target's behavior. This is certainly an area in which computational modeling can provide us a testbed for exploring different possibilities -- perhaps even serving as a source of potential hypotheses for further psychological experimentation. Let us momentarily postpone discussion of the mechanisms involved with steps 5 and 6 in order to wrap up our high-level description of the model.

One of the unique features of the model presented in figure 1 is the box labeled **Inherit Backward**. In brief, this aspect of the model allows for the flow of information computed in the **Alternate World** back to the **Real World**. On a domain-general view of belief ascription, this makes sense. With Nichols and Stich (2003), I assume that the function of simulating alternate worlds stemmed from evolutionary pressure to plan for future contingencies. But if worlds are cut off from one another, there isn't a route for actions planned out in alternate worlds to find their way back to the real world for execution. Even if the simulation process serves a lower-level function than planning – perhaps one of expectation generation, there still needs to be a way for computed expectations to be

compared to reality once the latter catches up with the former. One way that these phenomena might be captured is to have a free-flow of information back and forth between worlds. I assume information computed in other worlds to be accessible to the real world provided the ascriber has no contravening real-world reasons (either explicit or implicit) to exclude it from consideration.

To make the last claim a bit clearer, take a paradigmatic case of pretend-play: that of a child making a mudpie. Children (of a certain age) putatively use their knowledge of pie-making while engaged in the pretense, while overtly applying reality-oriented constraints when needed. So in service of the pretense, children forward-inherit pie-making beliefs into an alternate world where mud is actually pie-filling and so on. After planning and executing a sequence of pie-making steps within the alternate world, motor-intentions are sent back down to the real-world for execution and monitoring. What we typically see is children play-acting the "baking," and "slicing" of the pie; but when it comes to eating the pie, they might bring the slab of mud near to their lips, open their mouth but not put the mud inside. They then proceed to chew on the non-existent pie in their mouths and swallow. Everything I mention is consistent with pie-making and consumption norms, including our real-world distaste for eating mud. Our real-world knowledge that mud isn't edible cancels an action sent down from the pretense world; in the language used by Goldman (2006), the action is *taken off-line* and suppressed. For a computational rendering of the prior example, see Bello (2012).

Recently generated empirical data is consistent with the backward inheritance functionality I mention above. Once perspectives are spontaneously generated, there are measurable effects at the level of task-related reaction time if the generated perspective differs from our own perspective on reality. For example, Surtees and Apperly (2012) ask subjects to count the number of dots on the walls of an enclosed room. The room also contains a humanoid avatar. In a subset of trials, the avatar sees just as many dots on the wall as the subject does. In the rest, there are differences in how many dots the subject and avatar see, respectively. This is due to the fact that the avatar is facing a specific wall in the room, while the subject has a God's-eye perspective on the entire room. In these discrepant-perspective conditions, subjects are slower to respond, apparently confounded by perspectival differences. However this experiment only involves differences in line-of-sight. A recent study by Kovács and colleagues suggests that this effect extends to the spontaneously-tracked false beliefs of other agents (Kovács et al. 2010). Subjects were shown a ball, initially in front of an occluder. Another agent is present in the stimulus. The subject is instructed to press a button as soon as they know where the ball is located at the end of stimulus presentation. In phase 1, the ball either rolls behind the occluder or completely out of view with the agent still present. In stage 2, the agent leaves the room. In stage 3, the ball either (1) returns from out of view and settles behind the occluder or (2) moves from behind the occluder out of view. In stage 4, the agent returns and the subject is prompted to press the button. One of the prior combinations involves the other agent falsely believing the ball to be behind the occluder, and the subject being ignorant of the ball's location since it has moved from behind the occluder to a location out-of-view. Now, we have a situation in which the subject is ignorant while the agent is knowledgeable (leaving aside issues of veracity). Rather than reflecting long reaction-times associated with button-pressing in the absence of knowing where the ball is, subject RTs were reported to be consistent with similar trials in which the other agent had a false belief while the subject was knowledgeable about the actual location of the ball. What all of this suggests is that (at least) simple beliefs involving line-of-sight and location are spontaneously tracked in the presence of other agents. Without contravening reality-oriented reasons, subjects seem to behave in accordance with information generated in these perspectives – all in line with the account of backward inheritance that I've given above.

3 Some Tentative Conclusions

The purpose of this exposition was to motivate a model of belief ascription at the level of cognitively plausible mechanisms. I began this discussion by examining how well a purely functionalist picture of the mind comports with human data. I've argued that this sort of purist account fails to individuate mental states due to its' relatively coarse level of description, though what I've claimed certainly doesn't rule out other formulations of functionalism. As another side-benefit, the algorithm that I've specified relies heavily on the notion of the **Agent Detector** as a precursor for ascription, allaying fears about Ned Block's "China Brain" thought-experiment. One might think that the version of functionalism that I've attacked is a fairly uninteresting straw-man; but given recent work on pretend beliefs (Goldman 2006), imaginary desires (Egan & Doggett 2008) and the relationship between believing and delusions (Bayne & Fernandez 2008) in the literature, I am fairly comfortable with the strategy I've pursued here. Of course, this spells trouble for psychological theories of mental-state ascription that traffic in domain-specific accounts of the attitudes (Leslie 1994; Carey 1985).

Instead, I've suggested that perhaps some of the complications involved with individuating mental states can be resolved if we dig a bit deeper and try to understand mental-state ascription through the complimentary lenses of cognition and computation. Many of the complicated cases that are not covered by functionalism expressed only at the level of mental-state terms are plausibly taken care of by the model I've described in this paper. Both belief ascription and pretense rely centrally on the ability to simulate alternate (and most especially counterfactual) states of affairs, accounting for numerous psychological results that indicate shared variance between the two abilities (Riggs & Peterson 2000, Drayton, S., Turley-Ames, K.J. & Guajardo, N. 2011), corresponding closely to the **Amend** step of the algorithm given earlier in the paper. The **Inherit Backward** feature of the model makes computations happening in the simulated **Alternate World** accessible to the **Real World**, plausibly explaining why fictions have impact on real-world behavior, how pretend-play is possible, and why joint action that depends on reasoning about the beliefs of teammates is so rapid in some situations. While much has been left underspecified in this particular exposition of the model, the upshot of what I've argued is that at the very bottom, mental states might be more about what we (mentally) do, and less about how we represent.

References

Langland-Hassan, P. (in press). Pretense, imagination, and belief: the single attitude theory. *Philosophical Studies*.

Nichols, S (2004). Imagining and believing: The promise of a single code. *Journal of Aesthetics and Art Criticism 62* (2):129-39.

Gendler, T. S. (2007). Self-deception as pretense. *Philosophical Perspectives*, 21: 231–258.

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.

Kovács, Á.M., Téglás, E. & Endress, A.D. (2010). The social sense: susceptibly to others' beliefs in human infants and adults. *Science*, *330*, 1830-1834.

Surtees, A. & Apperly, I.A. (2012). Egocentrism and automatic perspective-taking in children and adults. *Child Development*, 83 (2), 452–460.

Birch, S., & Bloom, P. (2007). The Curse of Knowledge in Reasoning About False Beliefs *Psychological Science, 18* (5), 382-386.

Keysar, B., Lin, S.& Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89, 25-41.

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.

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

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46, 551-556.

Nichols, S. & Stich, S. P. (2003). *Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford University Press.

Goldman, A.I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, Oxford: Oxford University Press.

Bello, P. (2012). Pretense and cognitive architecture. *Advances in Cognitive Systems, 2, 43-58.*

Bello, P. (submitted). Folk concepts and cognitive architecture part II: Mental simulation and dispositionalism about belief.

Egan. A. & Doggett, T. (2008). Wanting things you don't want. *Philosopher's Imprint*. 7(9), pp. 1-17.

Bayne, T. & Fernandez, J. (eds.) (2008). *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation.* Psychology Press.

Leslie, A.M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S.A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* (pp. 119-148). New York: Cambridge University Press.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press/Bradford Books.

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.

Drayton, S., Turley-Ames, K.J. & Guajardo, N. (2011). Counterfactual thinking and false belief: the role of executive function. *Journal of Experimental Child Psychology*, 108(3), pp.532-48.