Folk Concepts and Cognitive Architecture: Mental Simulation and Dispositionalism About Belief

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

Abstract

In recent years, there has been an uptick of interest in the philosophy of psychology concerning cases of belief ascription that seem *prima facie* ambiguous. A number of authors have claimed these to be genuine instances of in-between believing: situations in which it's neither right to ascribe the belief that *P*, nor is it right to ascribe the belief that $\neg P$ to a target agent (Gendler 2008a, 2008b; Gertler 2011). Following Schwitzgebel (2010), I find the existence of such cases to be unsurprising if we take beliefs to be partially constituted by collections of dispositions. This paper proposes a relationship between belief-ascription via mental simulation and dispositionalist accounts of believing. Specifically, I show how the process model of belief-ascription presented in (Bello 2013 submitted) and a corresponding computational implementation is able to naturally accommodate situations in which it isn't quite clear what a target agent believes.

1 Introduction

Coming to an adequate account of the semantics of believing has been a longstanding aim for philosophers of mind, philosophers of psychology, cognitive scientists and researchers in Artificial Intelligence. But despite the best efforts of some very bright people, getting a handle on what it means to believe continues to present us with challenges. Indulging in a bit of introspection, we can see the kind of complexity that any reasonable model of believing must somehow manage to capture: we can believe things truly and falsely. We can be ignorant of certain facts, and deeply or weakly committed to the truth or falsity of others. We can be genuinely uncertain about certain propositions or about our beliefs themselves. We can ascribe beliefs to ourselves and to others – and these ascriptions are often full of complications. The complication that will occupy our discussion involves in-between believing, where it isn't entirely right to claim that a target agent fully believes either *P* or $\neg P$. For example, the following snippets are taken 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's impossible to constantly maintain such self-conscious vigilance […]."

What seems to be happening here, in essence, is a splintering of Juliet's professed beliefs away from the beliefs we would ascribe to her based on observing her behavior. One way to think about Juliet's beliefs about the races is in terms of her *dispositional profile*. On certain liberal accounts of dispositionalism such as the one defended in Schwitzgebel (2002), an agent believes *P* iff under favorable circumstances *C*, she acts, thinks and feels in roughly *P-ish* ways. In other words, the profile of the agent's thoughts, feelings and actions are *P-consistent*. So upon an ascriber observing enough of Juliet's overtly racist behavior toward African-Americans, the ascription "Juliet thinks that the races aren't of equal intelligence" may be justifiably made with respect to some context-sensitive definition of "enough." Similarly, Juliet's sincere utterance that the races are of equal intelligence counts as strong evidence of a proposition that she truly believes to be the case, since when *P* is truly believed in favorable circumstances she would presumably be disposed to assert that *P*.

Dispositions are tied closely to prediction and explanation because they inherently take the form of statements such as "If condition *C* holds then subject *A* will or may enter state *S*." Upon observing the conjunction *A* and *S*, we have some grounds for explaining *S* by appeal to *C* being the case. We also have grounds to predict that *A* will be in state *S* when we observe conditions *C*, *ceteris paribus*. Dispositions are modal generalizations in the sense that they scope over both actual instances of *C* and cases in which we say *A* would be apt to be in *S* were it the case that *C*. Following this line, I take a disposition to be both an actual or counterfactual conditional; and a dispositional profile associated with *P* to be the set of such conditionals that are *P-consistent*.

I claim that belief ascription via a properly nuanced version of mental simulation captures their dispositional profiles. After establishing the correspondence between mental simulation and beliefs as dispositional profiles, I show how the architectural machinery that implements the model of mental simulation described in Bello (2013 submitted) captures apparently ambiguous beliefs. Finally, I conclude with some brief words on how this particular account of beliefs as dispositions might be applied to other folk-domains of interest with an emphasis on moral judgments.

2 Mental Simulation and Dispositional Belief

The literature on the mechanism by which beliefs are ascribed typically turns on a distinction between ascription-by-simulation, and ascriptions made as a result of theory-laden inference (Davies & Stone 1995). Many, if not most researchers agree that ascription involves a combination of both approaches, although debates remain as to how the two interact and whether one collapses into the other in the limit (Nichols & Stich 2003; Carruthers 1996; Davies & Stone 2001). In the interest of brevity, I do not attempt a full-throated defense of my preferred approach to belief ascription with respect to collapse arguments in this paper. It is sufficient to say that I endorse a position in which mental simulation begets theory-laden inference, rather than the two being somehow in competition or operating largely in parallel.

Turning back to our prior discussion on dispositions, I reiterated the commonly held view that a disposition *d* is a modal generalization of a conditional taking the form: "(typically) if *C* then *A* will

(likely) *S*." While there are a variety of expositions available, almost every account of how beliefs are ascribed via mental simulation involves the ascribing agent doing something like the following:

1. creating a "mental space" that roughly corresponds to the target agent's mind
2. figuring out a set of candidate mental states the target agent might be enjoying,
3. creating "pretend" versions of these states,
4. populating the space with these "pretend" states,
5. utilizing the ascribing agent's practical reasoning system on the contents of the space
6. taking the results of step 5 offline

In cases where the ascribing agent believes *P* and knows that the target agent believes $\neg P$, the ascribing agent must suppress her egocentric perspective in order to maintain the distinction between what she knows, and her model of what the target agent knows. Inferences by way of simulation cash out to statements of the form "If I were *A* in circumstances *C*, I'd (likely) *S*." However we want to construe mental simulation, it seems clear that the upper-bounds of the kind of inference involved looks conspicuously like counterfactual reasoning. Simulation, when described this way gives a straightforward story of how dispositions might be manifested as I've described them. In cases where I actually *am A* in circumstances *C*, simulation describes a set of actual conditionals and partially maps to introspection. In cases where *A* is another agent assumed to be different from myself, simulation embodies a set of counterfactual conditionals corresponding to *A*'s dispositional profile[*].

Depending on what sort of information about *A* is known, simulation can capture everything from narrowly-scoped concrete dispositions to wider-scoped sets of dispositions associated with describing *A* in terms of personality traits. An instance of the former might be a disposition to scratch a particular itch on the leg once you notice that it itches, while the latter might describe a large set of plausible actions to be taken by *A* in circumstance *C* if *A* is deemed to be conscientious. To get clearer on how this works, I now turn to briefly reviewing a process-model of belief-ascription that I call CMBA, standing for *A Cognitive Model of Belief Ascription*. Once reviewed, I briefly sketch out how CMBA is implemented computationally, and show how Juliet's ambiguous beliefs can be captured.

3 CMBA: Overview and Implementation

In series of related papers, I have offered a hybrid account of belief-ascription driven largely by mental simulation and motivated by a commitment to domain-general cognitive capacities (Cassimatis 2006; Bello 2013 submitted). Figure 1 represents an abstract description of the process by which belief ascription occurs within CMBA. The basic algorithm-sketch 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.

[*] See Bello & Guarini (2010) for some earlier thoughts on 1st and 3rd-person ascription via simulation in a computational cognitive architecture.

7. **Supress**: Send inferred information in *w* back to the real world. If in conflict with anything currently known, amend and/or suppress.

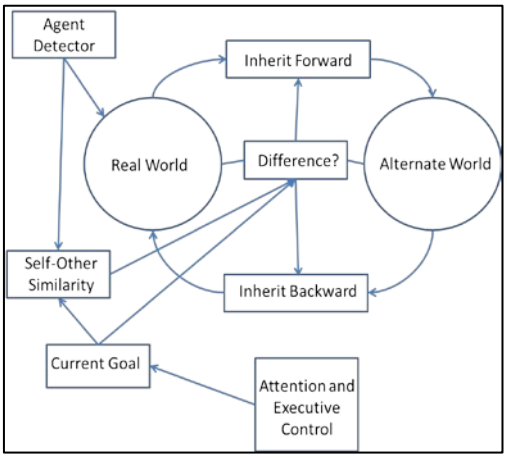

Figure 1: CMBA – Flow Diagram

The remainder of this section will discuss how this process model, defended in Bello (2013 submitted) is implemented computationally in the *Polyscheme* cognitive architecture (Cassimatis et al 2010).

3.1 Implementing CMBA in Polyscheme

The Polyscheme architecture is predicated upon the *Cognitive Substrate Hypothesis* (Cassimatis 2006), which roughly states that in almost every instance, higher-order cognition supervenes on a relatively small collection of domain-general representational and inferential resources. For the most part, these domain-general resources have been assumed to be those necessary for navigating the physical world, with minor additions when demanded by the relevant human data. The Polyscheme research program has largely consisted of deriving mappings between higher-order cognitive phenomena and the representational and inferential resources assumed by the cognitive substrate mentioned above. It has been shown that Polyscheme's current set of assumptions about the cognitive substrate are sufficient for capturing a wide variety of inferential strategies (deduction, abduction, induction), unification-based grammars for syntax parsing, aspects of linguistic semantics, pretense, and a variety of phenomena in belief ascription (Cassimatis et al. 2009a, 2009b; Uchida et al. 2012; Cassimatis 2004, 2009; Bello 2012; Bello et. al. 2007).

For our purposes, we take an instance of Polyscheme (and thus an instance of CMBA, shown in figure 2) to consist of the following: A set of *processing elements* corresponding to its set of domain-general inferential resources, a *cognitive focus of attention* that corresponds to what piece of information is currently being attended to, a *focus manager* that integrates inferences made by individual processing elements and broadcasts results back to them, and a symbolic *interlingua* through which the processing elements communicate with the focus manager. Only one bit of information can be attended to at a time, ensuring fine-grained integration of perception, cognition and action. I now turn to describing the interlingua, the internal representations used by the *Constraint* processing element and their related semantics.

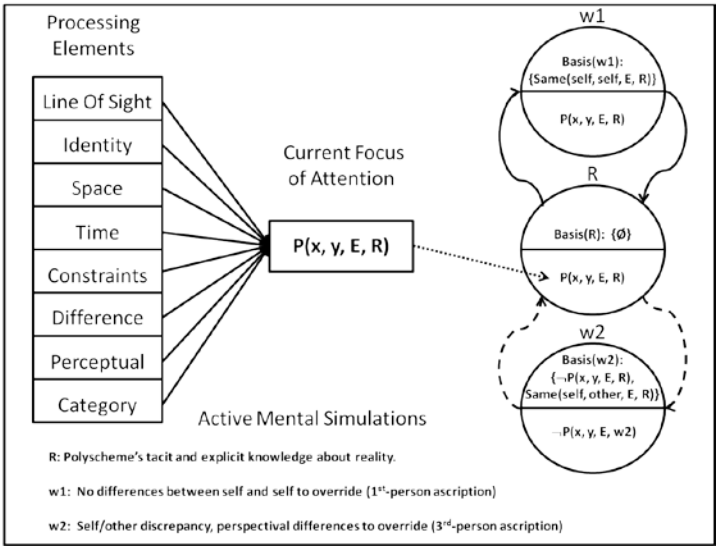

Figure 2: CMBA as Implemented in Polyscheme

3.2 Atoms and Constraints: Syntax and Semantics

While Polyscheme is designed to be an architecture that integrates inferences from multiple data structures and algorithms with differing representational facilities, all processing elements in the architecture cast the results of their computations into a logic-like interlingua. The basic units of knowledge representation in the architecture are called atoms. Atoms are relational elements having the form: *AtomName(arg1, …, argm, time, worldname)*, so to say that its sunny in DC all the time (in the real world), one might use the expression Sunny(dc, E, R), where E denotes "eternally" and R denotes the real world. This kind of representation can also be used to talk about finite temporal intervals and other possible worlds. If we wanted to say that it's sunny tomorrow in DC, we might express this as Sunny(dc, tomorrow, R), or if we wanted to say that it may be sunny in DC next week we say Sunny(dc, nextWeek, w). In the prior case, we replace the eternal interval E with a finite interval called tomorrow, and the atom expresses certitude that it will be sunny in DC tomorrow. In the latter case, we replace both the temporal interval and the world in which the atom has a truth value, signifying its expression of a mere possibility (relative to R)[†]. Standard logical conjunction and negation can be used to compose atoms in the usual way, yielding the expected sorts of expressions: *Atom1(arg1_1, …, argm_1, time_1, worldname_1)* $\wedge$ $\neg$*Atom2(arg1_2, …, argm_2, time_2, worldname_2)*.

In addition to standard logical notation for conjunction and negation, Polyscheme provides resources for representing two types of (implicitly) universally quantified *constraints* that can exist between collections of atoms. The first of these looks like the standard material conditional $\rightarrow$, and represents a *hard constraint* that quantifies over all worlds. To say that Sunny(?place, ?time, ?w) $\wedge$ In(?agent, ?place, ?time, ?w) $\rightarrow$ $\neg$Need(?agent, umbrella, ?time, ?w) is to say that for all worlds, if it

[†] Logical possibility as defined by the alethic modality is not treated here, although clever combinations of the mechanisms described in this paper along with the account of quantification given in Uchida et. al. (2012) can be utilized to cover possibility and necessity under a translation to FOL.

is sunny somewhere (represented by the variable ?place) at some time, and ?agent is an agent, then he won't need an umbrella at that time. But not all constraints are hard constraints. Polyscheme can express *soft constraints* using a numerical notation attached to the conditional operator of the following form: (*num*)→, where *num* assumes a value either in the range (0,∞) when performing weighted constraint satisfaction or in the range (0,1) when performing maximum likelihood calculations over worlds. For the sake of simplicity, I focus only on the former. To return to a slightly modified version of our previous example now expressed as a soft constraint, we have: ¬Sunny(?place, ?time, ?w) ∧ In(?agent, ?place, ?time, ?w) (.5)→ ¬Need(?agent, umbrella, ?time, ?w), saying that if in some world(s) at some time and place an agent is present and it's not sunny, then he won't need an umbrella. This allows us to explore worlds in which (1) it's not sunny and the agent doesn't need an umbrella and (2) worlds where it's sunny and the agent needs an umbrella (for whatever reason). Every time we encounter situations like (2), the world at which the atom has its truth value is penalized by .5. The *best world* consists of a set of ground literals (atoms with all variables bound and with truth assignments) that is both consistent and generates the least cost. Violations of hard constraints result in infinite cost. Since constraints are implicitly universally quantified and all atoms have world arguments we see that Polyscheme constraints implement a restricted form of modal generalization: just the kind of generalization useful for partially expressing dispositions. As I momentarily show, Polyscheme also provides a means by which to reason counterfactually using the same constraints, providing the rest of the resources for capturing dispositions applying to non-actual states of affairs.

Atoms have a distinctly different kind of semantics than that typically employed by formal logics or probabilistic calculi. Polyscheme implements a multivalent evidence-tuple associated with each atom. Evidence tuples consist of two entries: evidence for and evidence against a particular atom. As currently implemented, evidence can take on a number of values: **C**ertain, (Very) **L**ikely, **l**ikely, **m**aybe and **n**eutral. So Sunny(dc, tomorrow, R) <C, n> means that its certainly true that it's sunny tomorrow in DC in the real world, while Needs(me, umbrella, tomorrow, R) <n, L> means that it's highly unlikely that I'll need an umbrella tomorrow in the real world. Evidence tuples for atoms interact with the types of constraints mentioned above. If for example, I have a constraint Sunny(?place, ?time, ?w) ∧ In(?agent, ?place, ?time, ?w) → ¬Need(?agent, umbrella, ?time, ?w), and I know that Sunny(dc, now, R) and In(me, DC, now, R), then I will infer ¬Need(me, umbrella, now, R) <C, n> as I would with a standard application of *modus ponens*. However, if I know with certainty that In(me, dc, now, R) and I think it very likely that Sunny(dc, now, R), I infer ¬Need(me, umbrella, now, R) with <L, n>. Soft constraints propagate uncertainty associated with the antecedents of constraints to their conclusions. This feature of Polyscheme's semantics is crucial for implementing the kind of counterfactual reasoning needed to represent beliefs as dispositions.

Polyscheme allows for two distinct patterns of inference. The first pattern is what you might normally expect: one that matches the antecedents of constraints and infers consequences. The second pattern of inference corresponds to explanation: when a consequent is observed, Polyscheme attempts to infer the antecedents. This functionality is toggled by an internal variable called **shouldExplain** that can be set to true when explanation is required but is set to false as a default. Ascription-by-simulation uses both patterns of inference, the first for prediction and the second for explanation of an observed action taken by another agent. Finally, it should be noted that Polyscheme runs until quiescence. This means that wherever possible, atoms having anything but <C, n> or <n, C> as truth values are focused on and re-evaluated. The re-evaluation process is described in Bello (2012), and consists in simulating two worlds, one in which the uncertain atom is assumed to be true, and one where it is assumed to be false. If either world generates a contradiction, Polyscheme resolves the truth-value of the uncertain atom to the value assumed for the atom in the non-contradictory world.

3.3 Worlds, Inheritance and Inference

As was detailed in the last few paragraphs and illustrated visually in figure 2, Polyscheme atoms have their truth-values in *worlds*. To qualify, worlds in Polyscheme are not the maximally consistent states of affairs describing the totality of a universe of discourse as they are typically taken to be in Kripke-style possible worlds semantics. Instead, worlds in Polyscheme are more naturally thought of as *epistemic alternatives*: partially specified descriptions of what is assumed in a particular context and the supporting information that is perceived, recalled, and/or focused on to support inference. Any world *w* is defined by its *basis*, which is the set of atoms assumed to be necessarily true within it. The real world R is defined as having an empty basis, suggesting that any atom could potentially be true or false with respect to reality. Hypotheticals, counterfactuals and other kinds of worlds are defined by their basis. Polyscheme can imagine a hypothetical world *h* where *AtomName(arg1, …, argm, time, parentworld)* is true by inserting the hypothesized atom into the basis of *h* and specifying the world to which the hypothesized atom is relevant, which is marked as *parentworld* in the prior expression. So if Polyscheme know nothing about the weather in R, and it wants to hypothesize that it is possibly sunny, it creates a world wSunny with Basis(wSunny) = {Sunny(now, R)}. Because R is specified in the hypothesized atom, wSunny is deemed to be *relevant* to its parent world R. Similarly, Polyscheme can simulate a counterfactual world *c* using exactly the same procedure with the caveat that at least one member of Basis(c) is the truth-functional negation of an atom in its parent world.

Once children are related to their parent worlds via the specification of their basis, Polyscheme is able to use a variety of *inheritance rules* that govern how the truth-values of atoms in parent worlds are passed on to their children and how the truth values of atoms specific to children are passed back to their parents. This process corresponds to the **Inherit Forward/Backward** mechanisms in figure 1, and the connections between the Polyscheme worlds illustrated in figure 2. The forward and backward inheritance process is explained in some detail by Scally et. al. (2012), but for the sake of exposition, I will attempt to very briefly summarize. When evaluating an atom *A* in hypothetical world *h* having a parent world *p*, Polyscheme will check the ancestry of *h* to see if *A* has a specified truth value in *p* or *p's* ancestor worlds. If it does, *A* is assigned the identified value, and if it doesn't, *A* is focused on and evaluated by Polyscheme's processing elements. Consequently when we hypothesize *A* in *h*, we expect nothing about *h* to contradict something already known to be true in *h*'s parents. This algorithm keeps evaluation to a minimum and enables Polyscheme to compactly represent worlds only in terms of differences with respect to their ancestry. This kind of inheritance strategy minimizes computation, maximizes storage and contributes to the kind of reactivity one hopes to achieve in a large-scale architecture for cognition. As we will see in a moment, this bit of functionality partially provides a natural means to talk about *default belief ascription*, or the strategy that allows belief-ascribers to assume that target agents share the same beliefs possessed by the ascriber.

In the case of simulating counterfactual worlds, the inheritance strategy detailed in the last section will not work. Recall that when simulating a counterfactual world *c*, one of the elements of Basis(*c*) must be the truth-functional negation of an atom in *c*'s parent world *p*. If Basis(c) = {¬*A*} and *A* is true in parent world *p*, then application of the basic inheritance strategy leads immediately to a contradiction and the subsequent termination of inference in *c*. Instead, Polyscheme detects when one or more basis elements of a world *c* are the truth-functional negation of atoms in *c's* parentage. Once detection is performed and an atom *A* is focused on in *c*, the same ancestry-checking procedure takes place with the caveat that if a value for *A* is found in *c's* lineage it is assigned a decremented value in *c*. As a result, the only atoms in counterfactual worlds that are absolutely true or false are elements in

its basis. Every inherited atom comes in as less-than-certain. Since Polyscheme attempts to resolve uncertainty whenever possible, a post-inference counterfactual world *c* ends up being the closest world to its parent *p* up to the differences specified between *c* and *p* in *c's* basis. For a worked-out example of this process at work in detail, see the model of pretense outlined in Bello (2012).

As truth-values are assigned to atoms in child worlds, Polyscheme sends the resulting evaluated atoms back to their parents, but decrements their truth values under all circumstances. If we think about simulating hypothetical futures, backward inheritance delivers atoms back to the parent world that correspond to possibilities consistent with what's known. This is especially relevant when the parent world is R and the application of hypothetical reasoning is planning future actions. Since I am concerned primarily with belief ascription, I put aside the case of hypothetical reasoning and aim to describe the unique features of backward inheritance in counterfactual cases. In all instances of counterfactual reasoning, atoms in a counterfactual world *c* inherit backward into their parent *p*. Some concern is warranted here since atoms derived against the backdrop of counterfactual assumptions may contradict something known to be the case in the parent *p*. Because atoms are inherited backward with decremented truth-values, any conflicts will be resolved immediately by way of evidence combination as described earlier.

There are a special class of counterfactual world defined by having an identity statement of the form *Same(agent1, agent2, time, parentworld*) in its basis. As shown in figure 2, *agent1=agent2* in cases of self-ascription of occurrent beliefs, or with different agents in the case of ascription of beliefs from one agent to another.[‡] I shall refer to these as *ascription-worlds*. Agents are assumed to be identical to themselves in reality as a matter of course. However, on the definition of mental simulation given in section 2, the dispositions embodied by mental simulation are qualified by "If *agent1* were *agent2*, then …" Getting back to our discussion of backward inheritance, any atom derived in an ascription-world inherits backward into its parent. Referring to figure 1, if Polyscheme is maintaining an explicit goal to reason about whether an agent *agent2* (including itself) believes *B*, it creates a counterfactual world where *Same(self, agent2, E, R)* is true, inherits information from R using the one of the inheritance methods described in this section (e.g. default inheritance if Same(*agent1, agent1*) or counterfactual inheritance if Same(*agent1, agent2*)). If *B* follows, an atom taking the form Bel(*agent2, B, arg1_B, …, argm_B, time,* R) is inherited back into R with a likely truth value in the case of differences between the ascriber and the target agents and with a certain truth value in the case of 1st-person ascription. This process is described in section 4 and is central to belief ascription via simulation.

Presaging the next section, the distinction between tacit and explicit beliefs held by Polyscheme ought to be made clear. Simply put, the atoms in R having a truth value comprise Polyscheme's latent set of beliefs about the real world. Some of these beliefs will take the form Bel(self, *B, …, time,* R). These are Polyscheme's explicit beliefs – contingent upon an explicit effort made at some point in time to ascribe beliefs as detailed at the end of the last paragraph. Atoms not otherwise prefixed by Bel are considered tacit beliefs. These atoms still match the antecedents of constraints are sufficient to guide both inference and action without being explicit. Using this distinction we are able to capture cases in which Bel(*agent, P*) and ¬*P*, indicating explicit belief that *P* alongside a tacit representation of ¬*P* that issue in decidedly ¬*P-ish* actions.

[‡] Some cases of introspection are defined by an ascribing agent reasoning about a past or future version of itself, sometimes called 2nd-person ascription. These situations are treated as 3rd-person ascriptions without the incorrigibility associated with 1st-person ascriptions of occurrent mental states. See Bello & Guarini (2010) for a justification and description of the architectural mechanisms supporting this distinction.

4 Representing Juliet's Splintered Mind

With all of the last section in mind, we now have the resources to represent and reason about Juliet's beliefs given her sincere egalitarianism in thought and bigotry in deed. Let *S* be one of Juliet's male African-American students, and let *J* be Juliet. *S* will take the form of a CMBA-instance implemented in Polyscheme. What follows is merely a model-sketch, for the sake of brevity – it is in no way a complete or even an interesting description of what an average student might know about racist tendencies. Let *S* come equipped with the following set of constraints detailing the conditional dependencies between his tacit and implicit set of beliefs about racism:

1. Bel(?ag, RacesEqual, ?t, ?w) → VerballyAssert(?ag, RacesEqual, ?t, ?w)

2. IsA(?ag1, Caucasian, E, ?w) ∧ IsA(?ag2, AfricanAmerican, E, ?w) ∧ Prefers(?ag3, ?ag1, ?t, ?w) → ¬RacesEqual(?t, ?w)

3. HasChoice(?c, ?ag3, ?t, ?w) ∧ IsChoice(?ag1, ?c, ?t, ?w) ∧ IsChoice(?ag2, ?c, ?t, ?w) ∧ IsA(?c, CallsOn, E, ?w) ∧ Occurs(CallsOn, ?t, ?w) ∧ Agent(CallsOn, ?ag3, E, ?w) ∧ Patient(CallsOn, ?ag1, ?t, ?w) → Prefers(?ag3, ?ag1, ?t, ?w)

For the sake of simplicity, I will unrealistically assume that Juliet verbalizes her belief that the races are equal and simultaneously calls on a Caucasian student to answer a hard question on the blackboard while looking directly at *S*. This avoids the complication of temporally extended belief ascription and the ensuing evolution of *S's* judgments about *J's* beliefs as she speaks and acts over time. Tracing beliefs over time in Polyscheme has been addressed in the context of modeling false-belief and related developmental tasks in Bello et. al. (2007), but will be left aside for future application of the framework I've described. Let us assume that *S* observes *J* verbally asserting that the races are equal while staring at him and calling upon a Caucasian student *P* at the present moment. *S*'s observations correspond to the following set of ground literals:

IsA(P, Caucasian, E, R)
IsA(self, AfricanAmerican, E, R)
IsChoice(P, choice, now, R)
IsChoice(self, choice, now, R)
IsA(choice, CallsOn, E, R)

VerballyAssert(J, RacesEqual, now, R)
Occurs(CallsOn, now, R)
Agent(CallsOn, J, E, R)
Patient(CallsOn, P, E, R)

Upon observing J's overt behavior, he simulates a counterfactual world *c* where Same(self, J, E, R) is true in the basis along with the atoms {VerballyAssert(J, RacesEqual, now, R), Occurs(CallsOn, now, R), Agent(CallsOn, J, E, R), Patient(CallsOn, P, E, R}. These specify observed actions taken by J. Observing actions toggle the **shouldExplain** flag in Polyscheme to true, allowing for matching constraints in either direction. The first grouping of ground literals inherits counterfactually into *c* with likely truth-values and match constraint number 3, resulting in an inference to Prefers(J, P, now,

c). This inference fires constraint #2, leading to ¬RacesEqual(now, c). Observing J verbally assert that the races are equal lead to matching the consequent of constraint #1 and a corresponding inference to Bel(J, RacesEqual, now, c). Backward inheritance from c into R results in the following:

Bel(J, RacesEqual, now, R) <L, n>
Bel(J, Bel, J, RacesEqual, now, R) <L,n>
Bel(J, ¬RacesEqual, now, R) <L, n> (alternatively: Bel(J, RacesEqual, now, R) <n, L>)

Evidence combination and resolution of uncertainty in R results in:

Bel(J, RacesEqual, now, R) <L, L>
Bel(J, Bel, J, RacesEqual, now, R) <C,n>

Both of the above atoms seem to accurately characterize the situation at hand. From *S's* perspective, *J's* beliefs are ambiguous, but what seems clear to him is that *J* believes (of herself) that she believes the races to be equal.

5 Conclusions

In this paper I have elaborated and defended a process-model and corresponding implementation of ascription-by-simulation. When cast in a particular way, simulation seems to provide a suitable means by which dispositions and dispositional profiles can be embodied as modal generalizations that scope over actual and counterfactual situations. The highly specific version of ascription-by-simulation offered herein captures a variety of situations in which an agent's professed beliefs seem to come apart from beliefs that might be ascribed to them upon observing their actions, as in the case of Juliet. While still somewhat preliminary, I suspect the framework elaborated in this paper is rich enough to support a wide variety of ascriptions that have evaded formalization using standard computational techniques. For example, it is unclear from Schwitzgebel's example that Juliet is at all uncertain about what she believes or holds anything like inconsistent beliefs. What she does seem to hold are contextualized beliefs that are perfect candidates to be represented by dispositional descriptions. It is perfectly reasonable to assume that Juliet might have tacitly racist beliefs of which she is yet unaware. Even if she is aware of her racist tendencies, she might not say in either inner or externalized speech "I am racist,." or "I am not sure if I'm a bigot." Instead, she might chide herself for her hypocritical behavior and be more mindful of her habits in the future. Along similar lines, it also makes no sense to assume that Juliet's student is uncertain about her racism since in certain situations she acts like one. Through *S's* eyes, *J* believes herself to be egalitarian, even though *S* infers that she might not be. This might further lead *S* toward directing *J's* attention to her own behavior in the hopes that she recognizes it for what it is.

The research agenda ahead of me is rather ambitious. Basic work on making the inheritance and difference-monitoring processes in figure 1 contingent on executive resources in Polyscheme might be sufficient for capturing the failure of Juliet to be attentive to her bad habits at all times. With further modifications, the dispositional framework given here lends itself to capturing self-deceptive behavior, ascriptions of hypocrisy, pragmatism, expressions of overconfidence, underconfidence and even of the second-order regulation of first-order mental states found in the work of Harry Frankfurt on free will. On an even more ambitious note, this sort of framework captures one of the core intuition behind various strands of virtue ethics; exemplifying professed virtues involves bringing doing into line with thinking (or professions of virtue). The future is wide open and there is much to

be done. I hope the few halting steps taken in this paper lead to a full-blown sprint toward truly elaborate cognitive models of mental state ascription.

References

Gendler, Tamar Szabó (2008a). Alief and belief, *Journal of Philosophy* 105, pp. 634-663.

Gendler, Tamar Szabó (2008b). Alief in action (and reaction), *Mind and Language* 23, pp. 552-585.

Gertler, Brie (2011). Self-knowledge and the transparency of belief, in *Self-Knowledge,* Anthony Hatzimoysis (ed.) . Oxford: Oxford.

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

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief, *Nous, 36,* 249-275.

Bello, P. (2013 submitted). Folk-concepts and cognitive architecture: how we believe

Davies, M. & Stone, T. (eds) (1995). *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell Publishers.

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

Carruthers, P. (1996). Simulation and self-knowledge: a defence of theory-theory. In P. Carruthers and P.K. Smith (eds.), *Theories of Theories of Mind.* Cambridge University Press, pp. 22-38.

Davies, M. & Stone, T. (2001). Mental simulation, tacit theory, and the threat of collapse. *Philosophical Topics* 29 (1-2):127-73.

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. L. (2006). A Cognitive Substrate for Human-Level Intelligence, *Artificial Intelligence Magazine*, vol. 27, pp. 45-55.

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

Cassimatis, N., Bugjaska, M., Dugas, S., Muruguesan, A., & Bello, P. (2010). An architecture for adaptive algorithmic hybrids. *IEEE: Systems, Man & Cybernetics Part B,* 40(3): 903-914.

Cassimatis, N.L., Murugesan, A. & Bignoli, P. (2009b). Inference with relational theories over infinite domains. In *Proceedings of the Florida Artificial Intelligence Research Society Conference*, pp. 21-26.

Cassimatis, N.L., Murugesan, A. & Bignoli, P. (2009a). Reasoning as simulation. *Cognitive Processing* 10(4): 343-353

Uchida, H., Cassimatis, N.L. & Scally, J.R. (2012). Perceptual simulations can be as expressive as first-order logic. *Cognitive Processing* 13(4): 361-369

Cassimatis, N. L. (2004). Grammatical processing using the mechanisms of physical inference. In *Proceedings of the Twentieth-Sixth Annual Conference of the Cognitive Science Society*. pp. 192–197. Chicago, IL.

Cassimatis, N.L. (2009). Flexible inference with structured knowledge through reasoned unification. *IEEE Intelligent Systems* 24(4): 59-67

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

Bello P., Bignoli, P. & Cassimatis N. (2007). Attention and association explain the emergence of reasoning about false beliefs in young children. In *Proceedings of the 8th International Conference on Cognitive Modeling*. pp. 169-174. Ann Arbor, MI.

Scally, J.R., Cassimatis, N.L., & Uchida, H. (2012). Worlds as a unifying element of knowledge representation. *Biologically Inspired Cognitive Architectures*, 1, 14-22.