Introspection and Mindreading as Mental Simulation

Paul Bello (paul.bello@navy.mil)
Office of Naval Research, 875 N. Randolph St.
Arlington, VA 22203 USA

Marcello Guarini (mguarini@uwindsor.ca)
University of Windsor, Dept. of Philosophy, 401 Sunset.
Windsor, Ontario Canada N9B 3P4

Abstract

We present a sketch of a computational account of the relationship between certain aspects of introspection with aspects of third-person ascription of mental states (mindreading). The theory we propose is developed in large part as a reaction to what we perceive to be a lack of precision in the literature and a lack of experimental techniques to properly inform the debate on the relationship between 1st and 3rd-person ascription. We first discuss the set of phenomenology associated with self-ascriptions and other-ascriptions before briefly mentioning patterns of deficits associated with each. We sketch the very beginnings of a theory of mindreading in both the 1st and 3rd person within a computational cognitive architecture having mental simulation as one of its core operations. The theory we develop provides computationally-grounded explanations that are compatible with both clinical data and the phenomenology of 1st-person attribution.

Keywords: Mental Simulation; Cognitive Architecture; Metacognition; Mindreading; Philosophy of Mind.

Introspection and Mindreading

The ability to predict and explain behavior, both self- and other-generated, is a defining feature of human intelligence and a crucial phenomenon to be accounted for at the process-level; especially for those of us interested in computational theories of cognitive architecture. One of the major constituents of this ability takes the form of being able to ascribe mental states in service of behavior prediction and/or explanation. We will refer to mental state ascription more colloquially as "mindreading." Typically, mindreading is mentioned as being related to predicting and explaining the behavior of others, but what of our ability to report on our own mental lives? This ability is generally termed introspection, and one important scientific task will be to clarify its relationship (or lack thereof) to mindreading.

After presenting some of the generally agreed-upon phenomenological features of introspection, we briefly summarize the theoretical options for the mindreading-introspection relationship and some of their immediate entailments. Finally, we present our own account of their relationship in terms of a computational cognitive architecture capable of both 1st and 3rd-person ascription via mental simulation.

Introspection: Phenomenology

Characterizing the nature of introspection has been one of the most active areas of epistemology and the philosophy of psychology. This being the case, many distinctions have been made in the process, as definitions of what it is to introspect become ever-more specialized. While some of these distinctions have arisen from a priori philosophical analysis, the advent of novel experimental procedures and the further development of neuroscience have added a substantial amount of data on introspection that is providing constraints on what our theories of self-ascription look like.

Even with its many distinctions, there seem to be a few phenomenological features that all parties agree to be related to, if not constitutive of introspection (Schwitzgebel 2010). While there is a minority who believe that either we have no mental states like beliefs to introspect or that self-attributions are only unconscious, automatic processes of self-interpretation (Carruthers 2009); the majority of others agree that humans have a window on their mental lives. Most philosophical work in the area has been dedicated to clarifying the role, function, and features of introspection.

Following the discussion in (Schwitzgebel 2010), what mostly seems to be agreed upon is that:

1. Introspection is about the mental/internal, and thus not about the non-mental/external.
2. Introspective judgments are accompanied by a strong sense of certainty, even stronger than judgments about other forms of sense data.
3. Introspective judgments are relatively direct in the sense that they occur directly without needing to be inferred from other supporting data, supporting a distinction between detecting versus reasoning about one's mental states.
4. Introspection occurs in the "specious present," comprised of a very short time period just before and just after the introspective act.
5. While effortful and non-automatic, introspective judgments about one's own mental life seem easier to produce and less prone to subjective feelings of uncertainty than judgments about the mental lives of others.

Whatever sort of theory we intend to develop ought to at least coarsely capture these features and preferably provide explanations for them in terms of computational mechanism.

Psychological and Clinical Data

In the case of mindreading, it's been long established that those on the autism spectrum have deficits associated with mindreading; especially in regard to appreciating the false beliefs of others when trying to predict or explain their behavior (Baron-Cohen 1995). The same subjects have trouble engaging in spontaneous pretence, both self-directed and with other children. Of course, a small percentage of those on the autism spectrum are high-functioning enough to pass typical tests of false belief understanding, and more advanced tests that probe second-order false belief understanding. Results as to performance of autistic subjects on introspective tasks have been somewhat mixed. Some data suggest that autistics are capable of self-report and robustly utilize self-ascriptions of beliefs, intentions, desires and the like (Nichols & Stich 2003) to describe how they feel at randomly cued intervals. On a more contrarian note, the number of subjects in these experiments are small (N less than 5) and consisted of extremely high-functioning patients, blunting some of the force of such a charitable interpretation. Other experimental results with autistic populations suggest serious deficits with introspective judgments as well as mindreading.

Those diagnosed with schizophrenia provide a second set of clinical data on both mindreading and introspection. Recently, large scale studies conducted by (Sprong 2007, Corcoran 2001) have suggested deficits in mindreading across different categories of schizophrenia. Schizophrenia has long been thought of as a characteristic deficit in introspection and self-monitoring, with delusions resulting from an inability to properly identify stimuli as being generated internally by the operations of the mind (e.g. inner speech, volitional imagery) or externally by other sources (Frith & Done 1988).

A third set of individuals consists of those with severe brain damage or those who have for some reason, required a commissurotomy, or severing of the main bundle of neural fibers connecting the right and left hemispheres of the brain. It has been reported that this subject pool demonstrates that the left hemisphere of the brain generates unconscious, automatic self-interpretations of the form we mentioned earlier (Gazzaniga 1967). Finally we have numerous psychological studies purporting to show healthy subjects having only the most tenuous grip on their inner lives. Perhaps most famous are the early studies of Nisbett and Wilson demonstrating subjects' lack of insight into the processes whereby they arrive at a decision (Nisbett & Wilson 1977). In this case, the subject falls prey to a particular form of automatically induced bias, but is asked for an explanation for why they chose as they did. It's unclear to us and apparently to Nisbett and Wilson as attested in their later writings (Wilson, 2002) that these results challenge the notion of introspection as traditionally conceived.

Prior Work

As we've mentioned, introspection and mindreading have been perennial topics in the philosophy of mind, and have now become important areas of study for psychologists and neuroscientists. While it isn't feasible to even topically review the prior work in the area, two sets of items are worth mention. The first of these concerns the lack of consensus on how to perform experiments to test claims about introspection, and subsequently how to interpret the results. Many of the studies performed have subject pools with N < 5, and rely on hermeneutical analyses of written reports by these subjects to draw conclusions (Hurlburt & Heavey 2006). The second claim, which relates in a way to the first, is that while purporting to explain the variety of phenomena we've mentioned so far, contemporary theories of introspection (Carruthers 2009, Nichols & Stich 2003) provide little more than box-and-arrow diagrams and verbal argumentation to support their favored position. Much of the verbal argumentation is aimed toward giving a convincing interpretation for the so-called data on introspection, which itself seems to defy consistent analysis, even by co-authors (Hurlburt & Schwiztgebel 2007)! Many of these theories endorse one form or another of the so-called theory-theory, simulation theory, or modular theory of mindreading. While space doesn't allow for detailed descriptions of the commitments made by each of the preceding options, we think it to be generally the case that each provides a set of constraints as to how computations underlying both introspection and mindreading might be made. In very broad strokes, theory-theory is committed to the existence of a body of theoretical knowledge about how beliefs, desires and other mental states stand in causal relation to one another to enable the prediction and explanation of behavior. Various strains of theory-theory have been proposed to underwrite both mindreading and introspection (Gopnik 1993). One way that theory-theory can be applied is inside a cognitive module, which is somewhat isolated from central cognition, and houses specific representational and processing resources dedicated solely to mindreading and introspection. Modules are generally thought to implement specific computational constraints on the variety and complexity of information allowed in and out of them, but different theorists have different takes on what these constraints are (Carruthers 2009, Leslie & Thaiss 1992). Finally, simulation theorists propose that we use our own mental states and inferential resources to construct mental simulations of ourselves-as-the-target, where the target is an agent whose behavior is to be predicted or explained (Goldman 2006). Current theorists have used these frameworks to define their particular notions of mindreading and introspection. Along with interpretation of clinical and other data, constraints generated by theory-theory and its' alternatives have led researchers to draw conclusions about whether or not these two abilities are served by different or identical computational mechanisms.

Imprecision

What seems so curious to us is why these theorists choose to commit to any of the frameworks we just mentioned in the last section. In essence, both simulation theory and modular theories of mindreading were developed as reactions to what are perceived implausibilities associated with theory-theory. For example, questions remain about what the contents of such a theory would be and how inference is performed efficiently using them. Classical questions from the artificial intelligence perspective regarding computation over such theories in dynamic environments (e.g. the frame problem, the relevance problem and their cousins) have never been addressed by the leading proponents of theory-theory. In addition theory-theory seems to commit to theories about the mental states of others, but also theories about how mental states are manipulated by inference procedures. Having detailed theories of the inferential tendencies of others seems to be a bit of an intellectual stretch for many. Similar questions about the structure and constraints that modules impose plague supporters of modular ideas about mindreading and introspection. The imprecision we describe poses not only a problem for a theory-laden interpretation process, but also for off-line simulation theorists (Goldman 2006) and some simulation-theory hybrids (Nichols & Stich 2003). In these cases, the mindreader selects a number of "pretend" beliefs, desires, and other relevant mental states and inserts them into their own practical decision-making system, taking the result "off-line;" meaning, any actions inferred in light of these pretend states are not actually sent to the motor system for execution as they would normally be for non-pretend inputs. While at least one of us (PB) is sympathetic to simulation, it isn't clear on any account of simulation how the pretend inputs are selected for simulation in the first place. All of these concerns serve to illustrate a more general point about theories of mindreading. In general, those who propose conceptual models for mindreading do so with an eye to philosophical issues or to empirical data without regard to how computations performed by these models might take place.

We feel that computational implementation provides at least a coarse guide to how feasible one option might be over another. Most computational models have been of the false belief task (Wimmer & Perner 1983). Examples from (Goodman et al. 2006), (Bello et al. 2007) and (Berthiaume 2008) almost completely cover the space, which is somewhat disappointing, given the many hundreds of false belief studies and associated variants that have been conducted since Wimmer and Perner's original experiment. While space doesn't allow for a detailed discussion, we now turn toward sketching an implementation of mindreading and introspection in a computational cognitive architecture that captures some of the general phenomenology and is sensitive to the constraints imposed by psychological and clinical studies.

Cognitive Architecture

Descriptions of the Polyscheme cognitive architecture in which we have conducted our modeling efforts can be found in (Cassimatis et al. 2009). A detailed account of the architecture and how coordination is achieved between its various elements can be found therein. For the sake of exposition, we only describe architectural features that are central to our account of the mindreading-introspection relationship.

Cognitive Architecture: Specification

Polyscheme is comprised of a number of *processing elements* (PE's) that communicate with one another via a *focus of attention* (FoA). Each PE maintains its own proprietary memory, data structures, algorithms for elaborating propositions, and internal knowledge representation that maps onto propositional form. Every PE is wrapped in an interface that allows two-way communication with the FoA through a propositional language. Choices of what PE's to include in the architectural specification are made through appeal to evolutionary, cognitive developmental, neuroscientific, and computational constraints. The PE's that serve our purposes in explaining mindreading are represented in figure 1 and include rule matching, categorization, gaze detection, difference detection, identity hypothesis generation/evaluation, temporal and spatial reasoners, and a perceptual buffer.

Strings of the form $P(x_0, ..., x_n, t, w)$ are called *propositions*. Simply stated, *P* is a relation (i.e. *Loves, Hates, Color, MotherOf*) over the set of objects x_i during the temporal interval *t* in a world *w*, which bears a truth value. We designate "E" as the temporal interval containing all other temporal intervals. A proposition's truth-value is a tuple <F, A> consisting of the positive evidence for (F) and negative evidence against (A) the proposition and a scalar valence. Evidence takes on one of the following values: F, A $\in$ {C, L, l, m, n} representing *certainly*, *very likely*, *likely*, *maybe*, and *unknown*.

Cognitive Architecture: Mindreading

Propositions in Polyscheme have truth-values in mentally simulated worlds. Polyscheme's "beliefs" that are derived from perceptual data or via inference exist as propositions that are true in "R" or the real world; however the architecture is also capable of entertaining counterfactual, past, future-hypothetical, and other forms of simulated worlds. Polyscheme's "beliefs" about the real world are propositions with "R" in the final argument slot. What we're really interested in is how Polyscheme is able to identify and reason about the beliefs of other agents, including reflection on its own beliefs. In past work, we have shown how 3rd-person ascription is reducible to a substrate of domain-general representational primitives and processing elements including mental simulation of counterfactual worlds, reasoning about identity, categories, and by applying conditional rules (Bello et al. 2007). While this surely sounds like quite a lot of mechanism, all of these abilities seem to be roughly in place by two years of age in typical human children, and none of them implies any commitment to innate modules or core theories. We do take mental simulation to be a critical operation for the ascription of beliefs, which according to our theory proceeds in the following way:

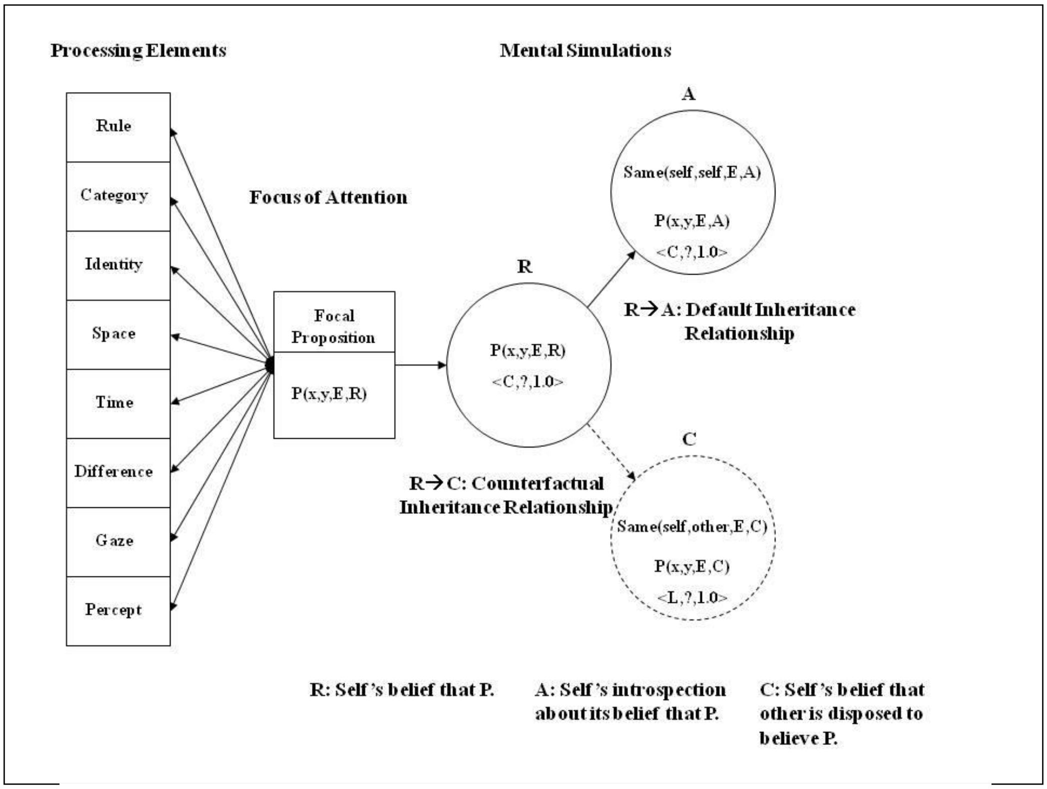

Figure 1: Polyscheme

1. Categorize other entity as an agent using category PE.
2. Construct counterfactual world C where Same(self, other, E, C) is true.
3. Detect differences between self and other using identity PE
4. Apply an override for each difference detected using conditional rule PE, forcing self-related propositions to resemble other-related propositions.
5. Proceed with inference and predict behavior appropriately.

The conditional rule PE implements a general-purpose rule that roughly looks like the following:

Holds(?P, self, ?t , ?w) ^ -Holds(?P, other, ?t, ?w) ^ Same(self other, E, ?w) ➔-Holds(?P, self, ?t, ?w)

Actual implementation of this rule is somewhat more complex, but incidental to our discussion. It suffices to say that mismatches between self and other-related propositions are detected as exceptions in simulated worlds C where Same(self,other,E,C) is true. An immediate concern is how such a rule fails to immediately generate a contradiction, since Holds(?P, self, ?t, ?w) is true, and –Holds(?P, self, ?t, ?w) is inferred as a consequent. Recall that propositions in Polyscheme have truth-values that are more differentiated than bivalent true or false. Also recall that Polyscheme's beliefs are propositions indexed to "R," the real world. Worlds in Polyscheme are related to one another via a process of *inheritance*. Inheritance relates a child world to a parent world, and operates in the following way: if during the course of inference, Polyscheme is asked to focus on a proposition P in a child world, it will check to see if P has a truth value in that world. If it doesn't, Polyscheme will look at the child's parent world to see if P has a truth value there. If it does, the truth value for P in the child world will be assigned the same value it has in the parent world. The inheritance procedure is visually depicted in figure 1 above. The inheritance procedure captures the idea that if we are to imagine a world in which some proposition like "pegasus exists" is true, other unrelated things we know about, such as "New York is north of DC" are vacuously true in our imagined world by virtue of the fact that they inherit truth values for these propositions from "R," the real world.

The rule we've given that performs an override looks like it might generate a contradiction. Polyscheme's world-simulation PE detects that Same(self,other, E, C) is a counterfactual claim, and when inheriting truth-values from the parent world “R” for propositions in the counterfactual child-world C, they inherit into C as only being very likely true or very likely false, rather than the certainly true or certainly false values they would be assigned if the counterfactual status of Same(self,other, E, C) was never detected. Since Holds(?P, self, ?t, C), etc. would inherit into C with less-than-certain truth values, Polyscheme can continue to infer in C without running into the danger of contradiction.

Inheritance, Overrides and Mindreading

How do inheritance and overrides in simulation relate to one another, and to both mindreading and introspection? We will differentiate between introspection of currently-held beliefs and 3rd-person ascription by appealing to different inheritance relationships with “R” that define them. Specifically, we are interested in the difference between *alternate* worlds and *counterfactual* worlds. We qualify what we mean by alternate world in the following fashion: an alternate world is such that no proposition in it is the truth-functional negation of a proposition in its parent world. For purposes of our discussion, “R” will always be the parent world of whatever simulations we are considering, whether they are alternate worlds or counterfactual worlds. This is in contrast to counterfactual worlds, which we’ve already explained, and which contain propositions that are truth-functional negations of propositions in their parent worlds. The difference between these two modes of simulation is illustrated in figure 1. When introspecting on currently-held beliefs, Polyscheme entertains an alternate world in which it is the same as itself. It does so by inheriting from its parent world “R” using an inheritance relationship called I_{aw}. We call this the “default” inheritance relationship since it perfectly preserves truth-values for propositions between parent and children worlds. In contrast, the counterfactual inheritance relationship, called I_{cw}, weakens the truth values for propositions inherited from a parent world R into a child world C, allowing counterfactual reasoning to proceed without immediately inferring a contradiction.

When introspecting, an alternate world A is considered in which Same(self, self, E, A) is true. According to the definition of strict identity, there are no differences between self and self, and thus nothing to override in such a world. However, when simulating oneself in the past or in the future, we might simulate a counterfactual world where Same(self, self_at_now-2, E, C) or a world where Same(self, self_at_now+10, E, C), and so on. Since these past or future versions of oneself might be importantly different from the standpoint of mental states, we note differences between these versions of ourselves and our current self, perform appropriate overrides, and make subsequent predictions or develop explanations. In this way, some sorts of introspective judgments work exactly the same way as 3rd-person ascription of mental states, while not committing us to the idea that introspection and mindreading are somehow identical and served by exactly the same set of cognitive operations (Carruthers 2009).

Accounting for the Data

Our theory satisfies a number of the conditions discussed in our introduction. Firstly, it should be clear that since we are simulating a world where we are ourselves, introspection about current mental states is clearly not aimed at perceptual features or external objects. The objects under consideration are propositions inherited from Polyscheme’s set of beliefs. This satisfies #1, the *mentality condition.* Since we differentiate simulating alternate worlds in which currently-held mental states are considered, versus counterfactual worlds in which either simulate ourselves as another agent entirely, or simulate ourselves in the past or future, there is a temporal constraint put on what we consider to be introspection proper. Simulation of past and future-selves certainly would count as self-knowledge, but there are acknowledged differences between self-knowledge broadly speaking, and introspection proper. This satisfies #3, or the *temporal locality condition.* Inheritance is not an inferential operation in the sense of having an associated logical operator with an associated semantics. Inheritance floats and attenuates the truth values of propositions from parent worlds to their children when required. In this way, truth of a proposition in a simulated world is arrived at non-inferentially, satisfying #3, the *directness condition.* Introspective judgments made in alternate worlds do not require any overrides relative to their counterparts arrived at counterfactually. If we associate some degree of effort or cognitive cost to performing an override of any sort, judgments about currently held beliefs will be guaranteed to seem at least as easy and likely much easier than judgments made about the mental lives of others, or of ourselves in the distant past or future. This satisfies the #5, the *ease condition.* Finally, properties of the two different inheritance relationships produce propositions in child worlds with different truth values. Inheriting from R into an alternate world produces propositions in the alternate world that have exactly the same truth value that they do in R. This contrasts to the relationship between propositions in R, and how they inherit into counterfactual worlds with slightly weakened truth values. This suggests that introspectively considered propositions are more certain than their non-introspective counterparts, satisfying #2, the *certainty condition.*

As for the clinical and psychological data, it’s difficult to speculate on how any existing model correctly accounts for disorders of mindreading and introspection. But speaking purely speculatively, some of the psychological data on confabulation (e.g. the Nisbett and Wilson results) can be attributed to the mechanisms in Polyscheme which produced its base set of beliefs in R. Since there is no requirement to have introspective access to the workings of these mechanisms, Polyscheme would merely take any propositional content generated by these mechanisms, and ascribe them to itself in an alternate world. In this way, Polyscheme has introspective access to the propositional content, without necessarily having access to the means by which it is acquired. In the case of autism, much has been said about cognitive deficits associated with autistic patients. Some of these deficits include the inability to follow and understand the targets of other agents gaze, thus eliminating a major source of evidence for understanding what other people currently believe. Other deficits have been hypothesized to include an inability to separate self versus other-centric representations, marked deficits in engaging in pretence and other forms of counterfactual simulation, and general lack of global coherence in cortical processing, all of which are critical elements of our story about mindreading and introspection. Similar deficits in schizophrenic subjects might be addressed by lesioning or confusing our inheritance and world-simulation mechanisms, which detect whether or not we're mindreading self or other-related targets. Of course, these are wild speculations, and we haven't produced any implementation. We only mention them to provide a prima facie story about how much deficits might be reproduced in a computational cognitive architecture.

Summary

We have given the rudiments of an account of the relationship between mindreading and introspection in an existing computational cognitive architecture using a single simulative mechanism, but having separate conditions of operation for each. We discussed our model's capacity to capture some of the defining features of introspection that have yet to be accounted for by competing models, providing a new way to generate and test hypotheses regarding the relationship between mindreading and introspection. While space hasn't permitted the inclusion of detailed computational models and associated model traces, these can be found for an example of 3rd-person ascription (the false belief task) and 1st-person ascription (the smarties task) on the first author's website: http://www.pbello.com/mindreading.html produced in a deprecated version of Polyscheme.

References

Schwitzgebel, E., (2010). "Introspection", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.).

Carruthers, P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*, 121-138.

Baron-Cohen, S, (1995). *Mindblindness: an essay on autism and theory of mind*. Cambridge MA: MIT Press.

Nichols, S. & Stich, S. (2003). *Mindreading: an integrated account of pretence, self -awareness, and understanding of other minds*, USA: Oxford University Press.

Sprong, M., Schothorst, P., Vos, E., Hox, J. & van Engeland, H. (2007). Theory of mind in schizophrenia: meta-analysis. British Journal of Psychiatry, 191(1), pp 5-13.

Corcoran, R (2001). Theory of Mind in Schizophrenia. In: D. Penn and P. Corrigan (eds.) Social Cognition in Schizophrenia. American Psychiatric Association, Washington DC.

Frith, C. & Done, C. (1988). Towards a neuropsychology of schizophrenia. *British Journal of. Psychiatry 153*: 437–43.

Gazzaniga, M.S. (1967). The split-brain in man. *Scientific American 217*, 24-29.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-259.

Wilson, Timothy (2002). *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge: Belknap Press.

Hurlburt, R. & Heavey, C. (2006). *Exploring inner experience*, Amsterdam: John Benjamins.

Hurlburt, R., & Schwitzgebel, E. (2007). *Describing inner experience? Proponent meets skeptic*, Cambridge, MA: MIT

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality, *Behavioral and Brain Sciences*, 16: 1–14.

Leslie, A.M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. Cognition, 43, 225–251.

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

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in children's understanding of deception. *Cognition*, *13*, 103–128.

Goodman, N. D., Bonawitz, E. B., Baker, C. L., Mansinghka, V. K, Gopnik, A., Wellman, H., Schulz, L. and Tenenbaum, J. B. (2006). Intuitive theories of mind: a rational approach to false belief. *In Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*.

Bello, P. Bignoli, P. & Cassimatis, N. (2007). Attention and Association Explain the Emergence of Reasoning About False Belief in Young Children. *In Proceedings of the 8th International Conference on Cognitive Modeling*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Berthiaume, V., Onishi, K. H., & Shultz, T. R. (2008) A computational developmental model of the implicit false belief task. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 825-830. Austin, TX: Cognitive Science Society.

Cassimatis, N., Bignoli, P., Bugajska, M., Dugas, S., Kurup, U. Murugesan, A. & Bello, P (2010). An Architecture for Adaptive Algorithmic Hybrids. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*.