

Simulation and the First-Person

Peter Carruthers

ABSTRACT: This article focuses on, and critiques, Goldman's view that third-person mind-reading is grounded in first-person introspection. It argues, on the contrary, that first-person awareness of propositional attitude events is always interpretative, resulting from us turning our mind-reading abilities upon ourselves.

Although Goldman (2006) advertises himself as defending a simulationist account of our knowledge of the minds of other people, it is important to note that he actually endorses a kind of simulation–theory mix. He concedes that there is a crucial role for theory at two different junctures in what he calls “high-level mind-reading”. One occurs whenever we wish to *predict* what someone in a given situation will think or do, in which case we must begin our simulation with some pretend inputs—but selection of the *right* inputs will need to be guided by theory. Likewise when trying to *explain* why someone has acted as she has, Goldman thinks that what we do is adopt a “generate and test” procedure—we try out some imagined inputs to the simulation process, and see if they result in an intention to perform an action of that sort. But since there are indefinitely many distinct inputs that we could in principle select and test, the choice of the most relevant and/or likely hypothesis will, again, have to be guided by theory.

In defending a mixed position, Goldman joins a number of other recent theorists who have argued likewise (Botterill and Carruthers, 1999; Nichols and Stich, 2003). Indeed, whether one thinks that mind-reading capacities are acquired via processes of hypothesis formation and testing (Wellman, 1990; Gopnik and Melzoff, 1997), or by the maturation of one or more innately structured component systems or “modules” (Baron-Cohen, 1995; Scholl and Leslie, 1999), one can agree with Goldman that processes of simulation play a crucial role in predicting and explaining the behavior of other people. What makes Goldman's view distinctive, and interestingly different from other mixed views on the market, is the foundational role that it accords to *introspection*, both in the course of mind-reading development in childhood and during mature adult mind-reading performance. He claims that first-person awareness of mental states is both prior to, and serves as the foundation for, our understanding the mental states of

others. In consequence he argues, in Chapter 9 of Goldman (2006), that self-ascription of mental states occurs via a process of introspective self-monitoring and classification that doesn't depend on theoretical knowledge. And then in Chapter 10 he argues that the core of our mental state concepts is constituted by an introspective code in the language of thought, which classifies our own internal states on the basis of their introspectible properties, again independently of theoretical knowledge. (These concepts will nevertheless be elaborated subsequently to contain such knowledge, of the sort thereafter utilized in third-person mind-reading.)

In the present article I shall subject Goldman's introspectionism to sustained critique. But in order to do that, it is important to have in place an appropriate foil. For these purposes I shall contrast Goldman's view with my own (Carruthers, 2006, forthcoming). According to the latter there is a single (albeit multi-component) mind-reading system, which is primarily *outward focused*. It evolved in the first instance for purposes of social understanding and engagement, just as proponents of "Machiavellian intelligence" and other similar views suggest (Byrne and Whiten, 1988, 1997; Dunbar, 2000). When we attribute mental states to ourselves we utilize the same conceptual and inferential resources that we use when interpreting others, with the result that our only access to a significant class of our own mental states is self-interpretative rather than introspective. Let me briefly elaborate.

Since the mind-reading system is outwardly focused, it needs to have access to perceptions of the environment. For in order to interpret the actions of others, it plainly requires access to perceptual representations of those actions. Indeed, I claim that, like most other conceptual systems, the mind-reading system can receive as input any sensory or quasi-sensory (e.g. imagistic or emotional) state that gets "globally broadcast" to all belief-forming, desire-forming, and decision-making systems. (For evidence supporting a global broadcasting architecture, see: Baars, 1988, 1997, 2002, 2003; Dehaene and Naccache, 2001; Dehaene et al., 2001, 2003; Baars et al., 2003; Kreiman et al., 2003.) As a result, the mind-reading system will find it trivially easy to self-attribute such states. If it receives as input a percept of a man bending over, for example, then it will be easy for it to form the belief, "I am seeing a man bending over." But I claim that while the mind-reading system has access to perceptual states, it has *no* access to the outputs of the belief-forming and decision-forming mechanisms that feed off those states. Hence self-attributions of propositional attitude events like judging and deciding are always the result of a swift (and unconscious) process of self-interpretation. However, it isn't just the

subject's overt behavior and physical circumstances that provide the basis for the interpretation. Data about perceptions, visual and auditory imagery (including sentences rehearsed in "inner speech"), patterns of attention, and emotional feelings can all be grist for the self-interpretative mill.

Like others who have written in defense of introspection for attitudes recently (Nichols and Stich, 2003), Goldman contrasts his view with a highly implausible form of "theory-theory", according to which self-attribution of judgments and decisions is grounded in observations of one's own behavior and circumstances alone. He writes:

That people use something like introspection can be made compelling by considering the implausibility of alternatives. I believe that I currently intend to walk into my study and remove a particular book from the shelf. What leads me to think that I have this intention? From what evidence could it be inferred—current behavior? past behavior? So far I have taken no step toward the study, so current behavior provides no clue. Nor do I have any past track record of taking that particular book off the shelf, so past behavior is no help either. The obvious explanation is that my intention-belief is obtained directly rather than inferentially. (2006, p.230.)

My reply to this argument is that Goldman's mind-reading system will have had access to a variety of forms of evidence in addition to overt behavior. He might, for example, have verbalized or partially verbalized his intention; or he might have formed a visual or proprioceptive image of himself selecting that particular book; or the context provided by his prior verbalized thoughts and visual images, together with a shift in his attention towards the door, might make it natural to interpret himself as having decided to walk to his study to collect that particular book. When Goldman's introspectionism is contrasted with an appropriate foil, therefore, it is by no means obviously superior.

In addition, there is now extensive evidence from cognitive science that people often lack introspective access to their own judgments and decisions, even in cases where they take themselves to have it. The evidence includes such facts as the following. First, split-brain subjects who are induced to perform an action by information presented only to their right hemisphere will nevertheless confabulate an explanation (using their left hemisphere) with all of the seeming introspective obviousness as usual (Gazzaniga, 1995). Second, normal subjects who are induced to make a movement via magnetic stimulation of motor cortex (but who are ignorant

of this fact) will claim to have been aware of *deciding* to make that movement (Brasil-Neto et al., 1992). And third, subjects' sense that they had intended an action, which was in fact performed by another person, can be manipulated by the simple expedient of having a semantically-relevant stimulus presented to them shortly before the action itself (Wegner and Wheatley, 1999).

Goldman is aware of, and acknowledges the force of, some of these data. In consequence he adopts what he calls a "dual-method theory" of self-knowledge, according to which we sometimes know of our own thoughts by introspection, and sometimes by self-interpretation. This is, indeed, consistent with the data mentioned above. One might propose, for example, that subjects only turn to self-interpretation when there exists no accessible judgment or decision, but where the circumstances strongly suggest to them that some such thought exists. Two points are worth making immediately, however. One is that the data show decisively that subjects are incapable of discriminating between introspection and self-interpretation on the basis of any subjectively-accessible cues. On the contrary, the confabulators described above interpret themselves with all of the same sense of introspective obviousness as normal. And the second is that a case can be made for thinking that a belief in the ubiquity of introspection might be built into the structure of the mind-reading faculty itself (either innately, or via "learning"), greatly simplifying its interpretative operations (Carruthers, 2008). So there is good reason to be suspicious of the common-sense intuition of the reality of introspection.

In any case, however, the dual-method account sketched above plainly won't work. For there are plenty of cases where perfectly ordinary judgments or decisions are actually present, but where subjects nevertheless confabulate. For example, provided that they no longer recall having been hypnotized, subjects who follow instructions given to them while under hypnosis will also confabulate explanations, while seeming to themselves to be introspecting (Edwards, 1965; Sheehan and Orne, 1968). Moreover, the social psychology literature on belief attribution is *rife* with studies demonstrating the effects of people's own behavior on the current judgments that they will mistakenly attribute to themselves (Eagly and Chaiken, 1993; Briñol and Petty 2003). Carruthers (forthcoming) reviews the patterning of these and other data and argues at length that the best explanation of it is that subjects never have introspective access to their own occurrent attitudes. I shall not repeat that argument here.

Goldman (2006, p.233) also argues *ad hominem* by quoting one of the cognitive scientists (Wilson, 2002) who has been most prominent in demonstrating the reality of confabulation,

showing that he actually seems comfortable with the idea of a “conscious mind”, to which we do have introspective access, existing alongside an unconscious one, to which we only have interpretative access. (Goldman could have found similar quotes in Wegner, 2002.) But Goldman here takes advantage of some careless writing. For there will of course be many conscious mental events (including visual imagery and inner speech) to which subjects have introspective access, and these events can often make a difference to subsequent behavior. But it is quite another matter to *identify* any such events with our occurrent judgments or decisions. On the contrary, when properly understood, they are always the effects or causes of such propositional attitude events (Carruthers, forthcoming). Hence one can allow that there is a “conscious mind” while also denying that we can ever introspect our own judgments and decisions.

Since mind-reading is grounded in introspection, on Goldman’s account, competence in attributing mental states to oneself should emerge in development prior to the ability to attribute such states to others. In contrast, since both self and other attributions are equally interpretative, on my own account, I predict no developmental difference in respect of self and other attributions of propositional attitude events. Goldman discusses, and briefly critiques, the arguments of Gopnik and Meltzoff (1994), who claim that the data suggest symmetry in the development of children’s capacities to attribute mental states to themselves and to others. He then introduces evidence of an asymmetry in his own predicted direction (self before other), expanding on the arguments of Nichols and Stich (2003). Let me comment briefly on the latter evidence.

Goldman mentions a study by Wimmer et al. (1988) which compared self and other understanding of knowledge versus ignorance in three year olds. In the *self* version, the children either looked into a box or didn’t, and were then asked whether or not they knew what was in the box. In the *other* version, the children observed someone else either look into the box or not, and were then asked whether that person knew, or didn’t know, what was there. The children did much better in the self version of this task, which Goldman takes to be evidence of his introspectionist position. But in fact the two tasks aren’t really comparable. For in the self version (but not the other version) the child can answer the question by accessing, or failing to access, knowledge of what is in the box. The child can, as it were, ask herself the first-order question, “What is in the box?”, answering the experimenter’s question positively if something comes to mind, negatively if not. Moreover, given recent data that children as young as fifteen

months can solve simple non-verbal versions of false-belief tasks (Onishi and Baillargeon, 2005; Southgate et al., 2007, Surian et al., 2007), it would be extremely surprising if three-year-olds lacked the underlying competence to reason about the ignorance of another person.

Goldman also appeals to studies that he interprets as showing that children's understanding of their own pretence emerges before their capacity to understand the pretence of others, again supporting his introspectionist position. But the experiments he appeals to aren't fully comparable. A study of children's understanding of third-person pretence (Rosen et al., 1997) is compared to a distinct set of experiments exploring children's understanding of first-person pretence (Gopnik and Slaughter, 1991). All sorts of different explanations of the divergent outcomes are therefore possible. Moreover, all of these tasks required children to offer verbal descriptions of the pretence in question, which might lead us to significantly underestimate their underlying competence. And indeed, recent non-verbal experiments demonstrate children's understanding of the pretence of another at 15 months, at or shortly before the age at which they first begin to engage in pretence for themselves (Bosco et al., 2006; Onishi et al., 2007).

The developmental data are neutral between Goldman's position and my own, therefore, or are actually rather more supportive of the latter. The other main area in which this battle can be fought concerns autism. Everyone agrees that third-person mind-reading is significantly impaired in autism, in which case my own prediction will be that autistic people's access to their own propositional attitude states must be impaired as well. Goldman's view, in contrast, is that introspection is intact in autism, with difficulties in other-understanding arising from difficulties in empathizing and perspective taking. Indeed, since many autistic people are especially good at the sort of focused learning and theorizing that gives rise to knowledge of the causal operations of complex systems, one would predict that this ability combined with introspective access to their own mental states should lead to them being especially good first-person mind-readers.

One set of data that Goldman discusses concerns an introspection sampling study conducted with three adult autistic men (Hurlburt et al., 1994; Frith and Happé, 1999). All three were able to report on what was passing through the minds at the time of a randomly generated "beep", although one of them experienced significant difficulties with the task. Goldman (2006, p.237) interprets this as demonstrating that introspection is intact in autism. I have two points to make. First, none of these three subjects was entirely deficient at third-person mind-reading. On

the contrary, two of them could pass second-level false-belief tasks, and the third could pass simple first-level false-belief tasks. So no one should predict that any of them would be entirely deficient at self-attribution, either. (It is worth noting, moreover, that the experimenters found a strong correlation between the subjects' abilities with third-person tasks and the sophistication and ease of their introspective reports.) Second, my own account predicts that autistic people should have no difficulty in reporting the occurrence of perceptual, imagistic, or emotional states, provided that they possess the requisite concepts. For these events will be globally broadcast and made directly accessible to their (damaged but partially functioning) mind-reading faculty. And indeed, much of the content of the introspective reports of the three subjects concerned visual imagery and emotional feelings. Reports of their own occurrent attitudes tended to be generic ("I was thinking ..."), and one of the three men (the one who could only pass first-level false-belief tasks) had significant difficulties in reporting his own attitudes at all.

Another set of data of the same general sort concern the autobiographical reports of autistic adults, who are often able to describe with some vividness what their mental lives were like at ages when they almost certainly wouldn't have been capable of attributing mental states to other people. Nichols and Stich (2003) comment that (provided we accept the memory reports as accurate), the individuals in question must have had reliable introspective access to their own mental states prior to having any capacity for third-person mind-reading. But actually we have no reason at all to believe that memory is itself a second-order (meta-representational) process. When I observe an event, a first-order representation of that event may be stored in memory. When that memory is later activated, I shall describe it by saying that I remember *seeing* the event in question. But it doesn't at all follow that the original event involved any meta-representation of myself as seeing. Likewise for other sorts of memories, and other sorts of mental events. The fact that autistic adults give meta-representational reports of their mental lives as children doesn't show that autistic children are capable of meta-representing their own mental states.

The data from autistic people that Goldman considers don't support his introspectionist position against my own interpretational account, then. But there are other data that Goldman doesn't discuss, which suggest that autistic people are decidedly poor at attributing propositional attitudes to themselves. Let me describe just a couple of strands of evidence here.

Phillips et al. (1998) tested autistic children against learning-impaired controls (matched

for verbal mental age) on an intention reporting task. The children had to shoot a “ray gun” at some canisters in the hopes of obtaining the prizes contained within some of them. But the actual outcome (i.e. which canister fell down) was surreptitiously manipulated by the experimenters (in a way that even adults playing the game couldn’t detect). They were asked to select and announce which canister they were aiming at in advance (e.g. “The red one”), and the experimenter then placed a token of the same color next to the gun to help them remember. After learning whether they had obtained a prize, the children were asked, “Did you mean to hit that [e.g.] green one, or did you mean to hit the other [e.g.] red one?” The autistic children were much poorer than the controls at correctly identifying what they had intended to do in conditions where there was a discrepancy between intention and goal satisfaction. For example, if they didn’t “hit” the one they aimed at, but still got a prize, they were much more likely to say that the canister that fell was the one that they had *meant* to hit. (Russell and Hill, 2001, were unable to replicate these results; perhaps because their population of autistic children, although of lower average age, had higher average verbal IQs, suggesting that their autism was less severe.)

Likewise Kazak et al. (1997) presented autistic children with trials on which either they, or a third party, were allowed to look inside a box, or were not allowed to look inside a box. They were then asked whether they or the third party knew what was in the box, or were just guessing. The autistic children got many more of these questions wrong than did control groups. And importantly for our purposes, there was no advantage for answers to questions about the child’s own knowledge over answers to questions about the knowledge of the third party.

In conclusion: although Goldman is surely correct that simulations, of various sorts, play important roles in mind-reading, his distinctively introspectionist position is ill-motivated, and faces serious difficulties.

References

- Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Baars, B. (1997). *In the Theatre of Consciousness*. Oxford University Press.
- Baars, B. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Science*, 6, 47-52.
- Baars, B. (2003). How brain reveals mind: neuroimaging supports the central role of conscious experience. *Journal of Consciousness Studies*, 10, 100-114.

- Baars, B., Ramsøy, T., and Laureys, S. (2003). Brain, consciousness, and the observing self. *Trends in Neurosciences*, 26, 671-675.
- Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.
- Bosco, F., Friedman, O., and Leslie, A. (2006). Recognition of pretend and real actions in play by 1- and 2-year-olds: early success and why they fail. *Cognitive Development*, 21, 3-10.
- Botterill, G. and Carruthers, P. (1999). *The Philosophy of Psychology*. Cambridge University Press.
- Brasil-Neto, J., Pascual-Leone, A., Valls-Solé, J., Cohen, L., and Hallett, M. (1992). Focal transcranial magnetic stimulation and response bias in a forced choice task. *Journal of Neurology, Neurosurgery, and Psychiatry*, 55, 964-966.
- Briñol, P. and Petty, R. (2003). Overt head movements and persuasion: a self-validation analysis. *Journal of Personality and Social Psychology*, 84, 1123-1139.
- Byrne, R. and Whiten, A., eds. (1988). *Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.
- Byrne, R. and Whiten, A., eds. (1997). *Machiavellian Intelligence II: extensions and evaluations*. Cambridge University Press.
- Carruthers, P. (2006). *The Architecture of the Mind: massive modularity and the flexibility of thought*. Oxford University Press.
- Carruthers, P. (2008). Cartesian epistemology: is the theory of the self-transparent mind innate? *Journal of Consciousness Studies*, 15.
- Carruthers, P. (forthcoming). Introspection: divided and partly eliminated. *Philosophy and Phenomenological Research*.
- Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79, 1-37.
- Dehaene, S., Naccache, L., Cohen, L., Bihan, D., Mangin, J., Poline, J., and Riviere, D. (2001). Cerebral mechanisms of word priming and unconscious repetition masking. *Nature Neuroscience*, 4, 752-758.
- Dehaene, S., Sergent, C., and Changeux, J. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Science*, 100, 8520-8525.
- Dunbar, R. (2000). On the origin of the human mind. In P. Carruthers and A. Chamberlain (eds.),

- Evolution and the Human Mind*, Cambridge University Press.
- Eagly, A. and Chaiken, S. (1993). *The Psychology of Attitudes*. Harcourt Brace Jovanovich.
- Edwards, G. (1965). Post-hypnotic amnesia and post-hypnotic effect. *British Journal of Psychiatry*, 111, 316-325.
- Frith, U. and Happé, F. (1999). Theory of mind and self-consciousness: what is it like to be autistic? *Mind and Language*, 14, 1-22.
- Gazzaniga, M. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga (ed.), *The Cognitive Neurosciences*, MIT Press.
- Gopnik, A. and Meltzoff, A. (1994). Minds, bodies, and persons: young children's understanding of the self and others as reflected in imitation and theory of mind research. In S. Parker, R. Mitchell, and M. Boccia (eds.), *Self-Awareness in Animals and Humans*, Cambridge University Press.
- Gopnik, A. and Meltzoff, A. (1997). *Words, Thoughts, and Theories*. MIT Press.
- Gopnik, A. and Slaughter, V. (1991). Young children's understanding of changes in their mental states. *Child Development*, 62, 98-110.
- Hurlburt, R., Happé, F., and Frith, U. (1994). Sampling the form of inner experience in three adults with Asperger syndrome. *Psychological Medicine*, 24, 385-395.
- Kazak, S., Collis, G., and Lewis, V. (1997). Can young people with autism refer to knowledge states? Evidence from their understanding of "know" and "guess". *Journal of Child Psychology and Psychiatry*, 38, 1001-1009.
- Kreiman, G., Fried, I., and Koch, C. (2003). Single neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Science*, 99, 8378-8383.
- Nichols, S. and Stich, S. (2003). *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press.
- Onishi, K. and Baillargeon, R. (2005). Do 15-month-olds understand false beliefs? *Science*, 5719, 255-258.
- Onishi, K., Baillargeon, R., and Leslie, A. (2007). 15-month-old infants detect violations in pretend scenarios. *Acta Psychologica*, 124, 106-128.
- Phillips, W., Baron-Cohen, S., and Rutter, M. (1998). Understanding intention in normal development and in autism. *British Journal of Developmental Psychology*, 16, 337-348.

- Russell, J. and Hill, E. (2001). Action-monitoring and intention reporting in children with autism. *Journal of Child Psychology and Psychiatry*, 42, 317-328.
- Scholl, B. and Leslie, A. (1999). Modularity, development, and “theory of mind”. *Mind and Language*, 14, 131-55.
- Sheehan, P. and Orne, M. (1968). Some comments on the nature of post-hypnotic behavior. *Journal of Nervous and Mental Disease*, 146, 209-220.
- Southgate, V., Senju, A., and Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18, 587-592.
- Surian, L., Caldi, S., and Sperber, D. (2007). Attribution of beliefs by 13-month old infants. *Psychological Science*, 18, 580-586.
- Wegner, D. (2002). *The Illusion of Conscious Will*. MIT Press.
- Wegner, D. and Wheatley, T. (1999). Apparent mental causation: sources of the experience of the will. *American Psychologist*, 54, 480-491.
- Wellman, H. (1990). *The Child's Theory of Mind*. MIT Press.
- Wilson, T. (2002). *Strangers to Ourselves*. Harvard University Press.
- Wimmer, H., Hogrefe, G., and Perner, J. (1988). Children's understanding of informational access as a source of knowledge. *Child Development*, 59, 386-396.