Theory of Mind, Computational Tractability, and Mind Shaping

Tad Zawidzki
Philosophy, Mind-Brain-Evolution Cluster, Mind-Brain Institute
George Washington University

1. Introduction

Philosophers and psychologists have traditionally understood "theory of mind" as a human capacity for understanding and predicting human behavior based on the attribution of unobservable mental states, like beliefs and desires. The classical model views the human "mind reader" as a kind of scientist, formulating hypotheses about the unobservable causes of the behavior of her fellows, and then testing them through observation (Gopnik & Wellman 1995). Many argue that the attribution of unobservable, theoretical mental states increases the power of human social cognition over mere sensitivity to patterns of observable behavior, of the kind that characterizes the social cognition of most, if not all non-human animals (Tomasello & Call 1997). In this paper, I review some familiar problems with this view and suggest a novel strategy for dealing with them. In section 2, I explain why the timely and accurate attribution of mental states appears to be a computationally intractable task. In sections 3 and 4, I consider two standard models of human cognitive architecture aimed at mitigating problems of computational tractability: modularity and fast and frugal heuristics, respectively. I argue that these are unlikely to help in the case of theory of mind. In the final section, I show how "mind shaping" (Mameli 2001; Zawidzki 2008) – roughly, the practice of socializing individuals in ways that make human populations more homogeneous – can mitigate some of the problems raised in the earlier sections.

2. The Apparent Computational Intractability of "Mind Reading"

In the philosophical literature on theory of mind, and in much of the psychological literature, beliefs and desires are taken to be the central mental states required to make sense of behavior. The central "law" of so-called "folk psychology" is, roughly, the following: if an agent desires that P, and believes that not P unless she does Q, then the agent will desire to do Q. However, this and related laws must inevitably be qualified by potentially indefinite numbers of exceptions. In principle, any behavior is compatible with any finite set of mental states, given enough adjustments in other mental states. For example, just because someone *says* she supports Barack Obama, does not mean that she *believes* he is the best candidate, or that she will act to get him elected, etc. She might believe that John McCain is the best candidate, yet, at the same time, *desire* to conceal this fact from her interlocutors.

Philosophers call this the problem of "holism" (Morton 1996, 2003). Behavior is not correlated with finite sets of mental states. Rather, behavior is correlated with *whole systems* of indefinitely many mental states (thus the term "holism"). The holism problem jeopardizes the *timely* accuracy of any theory of mind based on the attribution of mental states like beliefs and desires. Human social cognition is extraordinarily powerful, yet, at the same time, extraordinarily efficient. We can often accomplish dramatic feats of interpersonal coordination in constantly shifting, dynamic social circumstances, where there does not appear to be enough time to

consider and rule out all possible hypotheses about our interactants' mental states. It seems unlikely that some kind of brute search through all possible sets of mental states compatible with behavioral cues emitted by our interactants can support such fluid socio-cognitive competence.

It is easy to underestimate the severity of this problem. After all, aren't people largely alike in their thinking? Of course, there are minor differences based on age, gender, and cultural background, but these pale in comparison with our overwhelming cognitive similarities. Think, for example, of all the beliefs that the vast majority of all human beings share. We all believe that the sun rises in the east every morning; that it is darker at night than during the day; that dogs are animals; that summer days tend to be warmer than winter days, and an infinite number of other similarly banal propositions. There is also considerable, though not as dramatic, overlap in desires. The vast majority of human beings prefer satiation to hunger, shelter to exposure to the elements, the (at least) occasional company of other human beings to permanent solitude. The holism problem allows for the *possibility* that mental states vary radically between people, even if the limited behavior to which interpreters have access is similar. However, this is arguably a mere logical possibility. If, in the actual world, there is substantial homogeneity in mental states, then the holism problem is mere philosophical paranoia, akin to global skepticism. Accurate attribution of beliefs and desires, and consequent behavioral prediction may not be *guaranteed*, but given contingent facts about human cognitive homogeneity, it seems less problematic than the holism problem implies.

Most prominent models of human theory of mind assume that human populations are cognitively homogeneous. Goldman (2006) defends simulation[1] as a mechanism for reliable mental state attribution largely on the grounds that human interpreters and their targets are likely to share most of their cognitive states. Nichols and Stich (2003) argue that the attribution of most beliefs to interpretive targets involves the default attribution of the interpreter's own beliefs, on the grounds that people share the vast majority of their beliefs. Other models of human theory of mind implicitly make similar assumptions. On one prominent model, interpreters are like scientific psychologists proposing and testing hypotheses about others' mental states (Gopnik & Wellman 1995). On another prominent model, such mind reading involves, instead, the automatic and largely unconscious operation of an encapsulated cognitive module (Leslie 1994; Fodor 1995). On either model, the assumption of cognitive homogeneity seems indispensible to computational tractability: it dramatically shrinks the space of possible hypotheses compatible with behavioral data.

Below, I raise some problems for the assumption of human cognitive homogeneity, but even if it is granted that most human beings share most of their beliefs and desires, this is still not enough to explain the computational tractability of human mind reading. The reason is that mind reading is useful only if it helps behavioral prediction; it must make a practical difference to our attempts to anticipate and coordinate with the behavior of our fellows. But massive overlap in *dispositional* beliefs and desires[2] is not enough to explain how this is possible; for, what matters to behavior is not a subject's dispositional cognitive states but, rather, her currently occurrent,

---

[1] The use of the interpreter's own cognitive states and processes to model those of her targets.

[2] Dispositional cognitive states are non-active cognitive states that one is disposed to have. They are distinguished from occurrent cognitive states, which are the active, behavior-guiding cognitive states an agent tokens at any given moment.

behavior-guiding cognitive states. And even if there is massive cognitive homogeneity among human populations, this surely characterizes only dispositional cognitive states. It is obvious that there are typically dramatic differences in the occurrent, behavior-guiding states active in different individuals in similar circumstances.

Although agents may have many of the same dispositional beliefs and desires, which of these become occurrent and behavior-guiding in a specific context depends on contingent matters like mood, recent cognitive biography, etc. So, even if an interpreter assumes that her target shares many of her dispositional beliefs and desires, unless she has amazingly detailed knowledge of her target's recent biography, she cannot determine which of these beliefs and desires are currently guiding her target's behavior. Unless she knows whether or not her target has recently had a fight with a loved one, or has not eaten all day, or recently received some very good news that put her in a good mood, or other biographical details, it does not matter that her target likely believes that winters are colder than summers, and other banal propositions. Two individuals with similar dispositional beliefs and desires, engaged in similar behavior, in similar circumstances, may nonetheless have very different beliefs and desires actively guiding their behavior. If the point of mental state attribution is behavioral prediction, then interpreters must determine which beliefs and desires of their targets are currently guiding their behavior. And massive cognitive homogeneity in dispositional cognitive states is not sufficient for this. Only if interpreters had unrealistic access to the detailed recent biographies of their targets could they have a chance of determining which of their cognitive states are likely to be active. Since this is typically not the case, it remains a mystery how effective mind reading can be accurate and computationally tractable.

3. Why a Theory of Mind *Module* Cannot Help

"Modularity" is the classic response to issues of computational tractability. If the human mind-brain deploys *encapsulated*, computational modules, with *pre-specified*, *domain-specific* databases, tractably searchable by *dedicated* processes, then, it is often claimed, problems of computational tractability can be avoided (Carruthers 2006). The idea is that, when confronting some domain-specific problem, e.g., predicting the behavior of another person, the human mind-brain need not search all information to which it has access. Such problems trigger activity in dedicated, domain-specific modules – in the case of the social domain, the theory of mind module – which are informationally isolated from other parts of the mind-brain, making the search for solutions exponentially more tractable. However, it is not clear that modularity can help in the case of social cognition. The reason is that, as Currie and Sterelny (2000) point out, *any* information might be relevant to the tasks of interpreting and predicting human behavior. They give the example of detective work: figuring out who committed a crime and for what reasons, and predicting the criminal's next move, is precisely the kind of problem that a dedicated, informationally encapsulated module cannot solve. The reason we like detective fiction is that there is no way of knowing, in advance, what sorts of information might be relevant to cracking a case. In principle, the number of asparagus spears left on a plate, or other such unlikely facts might be relevant to determining a suspect's mental state.

Fodor (1983) introduced the notion of computational/cognitive modules, contrasting them with "central systems", which, he argued, are responsible for most belief fixation in human beings.

Unlike modules, central systems are, according to Fodor, "isotropic", i.e., any information is potentially relevant to belief fixation. He focuses on examples from science, e.g., information about fluid behavior turned out to be relevant to fixing beliefs about the behavior of light in Nineteenth Century physics. Arguably, much everyday reasoning is similarly isotropic. There does not seem to be a way of pre-specifying what kinds of information might be relevant to selecting among products in a supermarket, or deciding whom to date, etc., in the way there would need to be were such problems tractable by encapsulated modules.

If Fodor is right that most human belief fixation is isotropic, and therefore a product of non-modular central systems, then a general case can be made against the modularity of theory of mind. For, the goal of theory of mind is to determine what beliefs are likely operative in another agent. But, if Fodor is right, the processes by which an agent fixes her beliefs are isotropic and therefore non-modular. So any interpreter of that agent cannot, herself, rely on some modular, informationally encapsulated theory of mind to determine which beliefs the agent will acquire and act on. If the interpretive target's decisions are determined by non-modular processes then so must be the interpreter's hypotheses about those decisions, if they have any chance of succeeding. Furthermore, besides figuring out how her target solves the belief fixation problems she faces, the interpreter must also determine to what information the target likely has access and what problems the target is most motivated by – problems the target need not solve. This compounds the problem facing the interpreter – not only must she, in effect, solve the same isotropic belief fixation problems as her target, she must also determine the parameters governing her target's solutions. Thus, the problems of computational tractability that arise for theory of mind do not appear to admit of a classical modularist solution.

4. Why Fast and Frugal Theory of Mind Heuristics Cannot Help Either

Fodorian modularity is an extreme solution to the problem of computational tractability, which seems unhelpful in many domains, particularly social cognition. However, there may be other kinds of modularity that evade some of the problems that have been raised for Fodorian modularity. Recently, Carruthers (2006) has defended a kind of modularity that is *not* based on some pre-specification of the *kind* of information that might be relevant to solving tasks in some domain. Instead, Carruthers argues that the problem of computational tractability can be solved using content-neutral, "fast and frugal heuristics" (Gigerenzer et al. 1999). Because such heuristics are content-neutral, there are no limits on the kinds of information they can consult. This is promising in the case of mind reading since, as we have seen, almost any kind of information can be relevant to this task. However, computational tractability is maintained by fast and frugal heuristics because there are strict limits on the *quantity* of information they can consult.

For example, "Take the Best" is a well-known fast and frugal heuristic. It requires that one recall criteria previously used to distinguish between alternatives in some domain, determine which criterion distinguished best, and use that criterion on one's current decision. For example, when asked which of two German cities is larger, one might recall that, previously, having a professional soccer team distinguished best between larger and smaller cities, and so one asks which, if either, has a professional soccer team. If neither or both do, one then proceeds to the next best criterion. In order to avoid intractable search, "Take the Best" has a "stopping rule"

that suspends search if it cannot arrive at an answer after some small, finite number of iterations (Gigerenzer et al. 1999; Carruthers 2006).

Fast and frugal heuristics combine computational tractability with openness to a wide variety of potentially relevant information. For example, although "Take the Best" is restricted to considering only information that a particular agent has recently consulted when reasoning about a certain domain, this restriction is content neutral. It includes different information for agents with different histories of reasoning about a domain. There is no reason why, for a different agent reasoning about relative population sizes of German cities, [3] "Take the Best" could not consult relative crime rates instead of presence of professional soccer teams (Carruthers 2006). Such heuristics are computationally tractable because they restrict search based on an agent's current epistemic context, including relevant recent searches, not because they restrict search based on the content of the relevant domain, e.g., social, or physical, etc. So there is no need to pre-specify the kinds of information likely to be relevant to each domain. This mitigates the problem that Currie and Sterelny (2000) raise for modular theory of mind. Any information, e.g., the number of asparagus spears left on a plate, may be relevant to a theory of mind task. But only information that has recently been useful on similar tasks is consulted on any particular occasion.

Unfortunately, this proposal seems unlikely to work for theory of mind tasks. The problem is that fast and frugal heuristics work only in domains characterized by extreme homogeneity. "Take the Best" solves new problems based on strategies that a particular agent has successfully applied to similar problems in the recent past. Unlike more sophisticated, statistical learning algorithms, such heuristics make no attempt to insure that the sample from which they generalize is unbiased. Their quickness and frugality consists precisely in the fact that they avoid such formal niceties. Strategies that *happen* to have worked for a particular agent in the recent past are taken to be appropriate for current and future problems. Such contingent regularities can safely be assumed in certain extremely homogeneous, well-behaved domains. However, there is every reason to deny that the social domain is like this.

Ironically, if human beings rely on fast and frugal heuristics to fix their beliefs, then the dependence of such heuristics on the idiosyncratic background knowledge of particular individuals is likely to make discovering their beliefs computationally intractable for fast and frugal *theory of mind* heuristics. This is because such social cognition would require quickly and frugally uncovering the idiosyncratic background knowledge on which one's interpretive targets rely. But fast and frugal theory of mind heuristics can depend only on what has worked for the *interpreter* in the recent past, and there is no reason to think that this is any guide to the idiosyncratic background knowledge of a new interpretive target. Furthermore, as Sterelny (2003) emphasizes, human beings appear to have strong, biological incentives to behave in ways that are unpredictable relative to heuristics that their potential competitors have previously used to predict them. Lastly, there are good reasons to think that individual variation among human beings is extreme, compared with other species: we have unmatched capacities for creative cognition and conation which involve random processes heavily dependent on idiosyncratic learning history (Carruthers 2006), and extreme phenotypic plasticity is likely a human

[3] Or for the same agent at a different time.

adaptation to extremely variable physical and social environments (Sterelny 2003). As Carruthers puts it,

> … the processes that produce creative thoughts may be partly random or chaotic [and] … highly sensitive to the particular beliefs and learning history of the agent, as well as to the specific cognitive context in which the creative thought is produced (such as the thoughts that had been active shortly before, the perceptual contents recently caused by the detailed environment of the agent, and so forth). (Ibid., pp. 288-289)

So, we have every reason to suppose that fast and frugal heuristics, the reliability of which relies on extreme homogeneity in the domains to which they apply, cannot support reliable social cognition.

Perhaps there are more specific mind-reading heuristics that can evade some of these problems. The most influential models of fast and frugal social cognition appeal to some kind of simulation (Goldman 2006). Interpreters save on the computational costs of interpretation by simply projecting, in some sense, their own decision procedures onto others. But the accuracy of such simulation heuristics obviously depends on extreme homogeneity in human populations: interpreters and their targets must prioritize problems in similar ways and make decisions based on similar information and heuristics. And, as we have seen, there are good reasons to doubt that such homogeneity exists in human populations.

5. Mind Shaping as Human Homogenizer

Let me end by proposing a sketch of how I think humans solve the problems reviewed above. I propose that certain low-level, automatic mind-shaping mechanisms, prevalent in human populations, work to homogenize them, thereby making fast and frugal theory of mind heuristics more effective.

Suppose human beings have an automatic, default disposition to compare their own behavior to that of others, monitoring for any discrepancies. If the other is higher status, discrepancies tend automatically to issue in attempts at self-modification: one tries to change one's dispositions such that one's behavior better matches that of the high status model. If the other is lower status, discrepancies tend automatically to issue in attempts to modify the other: one tries to change the other's dispositions, as in teaching offspring or punishing norm transgressors, such that the other's behavior better matches one's own. Assuming that judgments of status are largely homogeneous in a population, such mind-shaping dispositions would tend to further homogenize populations, counteracting the "centrifugal" forces causing individual variation. And such homogeneity would render fast and frugal theory of mind heuristics more reliable. For example, the "Take the Best" heuristic would be more likely to work. In a population of similarly socialized individuals, decision strategies that happen to have worked well in recent interpretive contexts are more likely to work in new circumstances. Similarly, variations of the simulation heuristic would also be more reliable: procedures that interpreters use in their own decision-making would be more likely to be used by their interpretive targets as well.

Solutions to problems of computational tractability that arise for theory of mind, I want to urge, lie not *within* human mind readers, but, rather, *outside* of them. Rather than deploy intractably sophisticated theories of each other's minds, we make use of a variety of low-level mind-shaping

dispositions to insure that our fellows are sufficiently familiar so that fast and frugal heuristics can help us accomplish our socio-cognitive goals. We teach our children to behave in ways that make them easier to interpret (Bruner 1983, Mameli 2001, McGeer 2001). We sanction those who behave in ways that are harder to interpret (think of the damage to status which often results from weakness of the will or absentmindedness) (Zawidzki 2008). We display unconscious, automatic, and irresistible tendencies to conformity, such as the "chameleon effect" (Chartrand & Bargh 1999). Such mechanisms shape the socio-cultural environment in ways that make coordination exponentially more tractable than it would be were interpreters to exhaustively search the mental-state-hypothesis spaces compatible with the finite sets of behaviors to which they realistically have access.

References

Bruner, J. 1983. *Child's talk*. New York: Norton.

Carruthers 2006. *The Architecture of the Mind*. New York: Oxford University Press.

Chartrand, T.L., and J.A. Bargh. 1999. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76: 893 – 910.

Currie, G., and K. Sterelny 2000. How to think about the modularity of mind reading. *Philosophical Quarterly* 50: 145-60.

Fodor, J. 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press.

_____. 1995. A theory of the child's theory of mind. In *Mental simulation*, edited by M.Davies, and T. Stone. Oxford: Blackwell.

Gigerenzer, G., P. Todd, and the ABC Research Group 1999. *Simple heuristics that make us smart*. New York: Oxford University Press.

Goldman, A. 2006. *Simulating minds*. Oxford: Oxford University Press.

Gopnik, A., and H. Wellman. 1995. Why the child's theory of mind really is a theory. In *Folk psychology*, ed. M. Davies and T. Stone, 232 – 58. Oxford: Blackwell.

Leslie, A. 1994. Pretending and believing: Issues in the theory of ToMM. *Cognition* 50: 211-238.

Mameli, M. 2001. Mindreading, mindshaping, and evolution. *Biology and Philosophy* 16: 597 – 628.

McGeer, V. 2001. Psycho-practice, psycho-theory and the contrastive case of autism. *Journal of Consciousness Studies* 8, no. 5 – 7: 109 – 32.

Morton, A. 1996. Folk psychology is not a predictive device. *Mind* 105: 119 – 37.

_____. 2003. *The importance of being understood*. London: Routledge.

Nichols, S., and S. Stich. 2003. *Mindreading*. Oxford: Oxford University Press.

Sterelny, K. 2003. *Thought in a Hostile World*. Oxford: Blackwell.

Tomasello, M., and J. Call. 1997. *Primate cognition*. New York: Oxford University Press.

Zawidzki, T. 2008. The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations* 11(3): 193 – 210