Simulation, Simplicity, and Selection:

An Evolutionary Perspective on High-Level Mindreading

Armin W. Schulz – University of Wisconsin-Madison

In this paper, I argue that a natural selection-based perspective gives reasons for thinking that the core of the ability to mindread cognitively complex mental states is subserved by a simulationist process – that is, that it relies on non-specialised mechanisms in the attributer's cognitive architecture whose primary function is the generation of her own decisions and inferences. In more detail, I try to establish three conclusions. Firstly, I try to make clearer what the dispute between simulationist and non-simulationist theories of mindreading fundamentally is about. Secondly, I try to make more precise an argument that is sometimes hinted at in support of the former: this 'argument from simplicity' suggests that, since natural selection disfavours building extra cognitive systems where this can be avoided, simulationist theories of mindreading are more in line with natural selection than their competitors. As stated, though, this argument overlooks the fact that building extra cognitive systems can also yield benefits: in particular, it can allow for the parallel processing of multiple problems and it makes for the existence of backups for important elements of the organism's mind. I therefore try to make this argument more precise by investigating whether these benefits also apply to the present case – and conclude negatively. My third aim in this paper is to use this discussion of mindreading as a means for exploring the promises and difficulties of evolutionary arguments in philosophy and psychology more generally.

Simulation, Simplicity, and Selection:

An Evolutionary Perspective on High-Level Mindreading

1. Introduction

The extent and quality of the human ability to read minds – that is, to attribute mental states to others and to predict their behaviour in terms of them – surpasses that of any other species on the planet. This sort of fact raises a number of questions; two of the more obvious ones among these are:

(1) What is the nature of the cognitive mechanisms underlying this ability?

(2) How did these mechanisms evolve?

These questions are logically independent from one another; however, this does not mean that an answer to (2) cannot also be useful as a means of getting closer to an answer to (1) (see also Sober and Wilson 1998). In this vein, I here appeal to *natural selection* to complete and make more precise an argument for a specific account of the nature of the mindreading mechanism. In particular, I try to show that natural selection supports the idea that 'simulationist' views mindreading – i.e. views that deny that mindreading requires specialised cognitive mechanisms – are more compelling than their competitors *since they are more parsimonious*.

The paper is structured as follows: in section 2, I review the most important current theories of mindreading, and make the dispute surrounding them more precise. In section 3, I develop and defend the evolutionary argument for a simulationist theory of mindreading. In section 4, I assess what contribution this argument can make to the debate at issue. I conclude in section 5.

2. Two Views of the Mindreading System

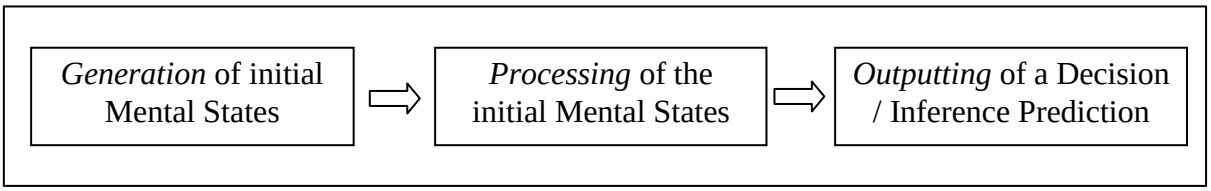
At present, there are two major types of theories of mindreading being defended (see e.g. Nichols and Stich 2003; Goldman 2006): the set of *information-rich* theories (comprising the 'theory-theory' and the 'modularity theory'), and the set of *information-poor* theories (comprising the 'simulation theory').[1] Intuitively, these two sets of theories are distinguished by the amount of extra 'stuff' (e.g. *information* about how our minds work, or specialised cognitive *mechanisms*) they claim needs to be added to a non-mindreading system in order to turn it into a mindreading system. Defenders of an information-rich view think that a lot of this extra stuff is needed (see e.g. Gopnik and Wellman 1994; Fodor 1992; Carruthers 1996, 28-31),[2] whereas defenders of an information-poor view think only very little extra stuff is needed, since many of the elements a non-mindreading system can easily be co-opted for mindreading purposes (see e.g. Goldman 2006; Nichols & Stich 2003).

Surprisingly, while much has been written in defence of either of these two views, the difference between them has actually been left quite vague. For this reason, the first step in the present inquiry into the nature of our mindreading system must consist in making this difference more precise. To do this, I start by determining the core functional elements of the mindreading system.

[1] Three things are worth noting about this classification schema. Firstly, there is a type of theory – the 'Rationality Theory' – that is neither information-rich, nor information-poor; however, since it is no longer seen to be a credible contender in this debate (see e.g. Nichols and Stich 2003, 142-148; Goldman 2006, chap. 3), it will not be considered further here. Secondly, the term 'theory-theory' is sometimes used synonymously with that of 'information-rich mindreading' (e.g. in Nichols and Stich 2003, 102); this, though, seems to be merely a terminological issue of little importance. Finally, all of these theories operate at Marr's 'computational' level of the cognitive hierarchy: they do not actually describe the *algorithms* used by the cognitive system or the neural *realisation* of these algorithms, but only the *problems* they have to solve (a point that will become important again below - see especially note 12); see also Nichols and Stich (2003, 10-11, 209-210).

[2] The 'theory-theory' and the 'modularity theory' are distinguished by their take on what this 'extra stuff' is (e.g. how the extra information being acquired, stored and processed, or what the nature of the specialised cognitive mechanisms is). Since this is not so important here, however, I shall not discuss it further.

When it comes to this, it is widely accepted that this system has two major components (Goldman 2006; Nichols and Stich 2003; Sterelny, 2003, chap. 11): on the one hand, we attribute mental states to someone else on the basis of various forms of mental and non-mental *evidence* (e.g. her perceptual states, her and others' linguistic and non-linguistic behaviour, the beliefs and desires *we* are holding, etc.); on the other, we attribute mental states to someone else on the basis of the theoretical and practical inferences we *predict* she is making. Call the first 'generation' and the second 'processing'. Note that these two parts are related: in order to predict a target's decision (in a mentalistic, non-behaviourist way), we need to use the individual, specific mental states that are the result of the generator as our starting points. Graphically, the situation can therefore be depicted as follows:[3]

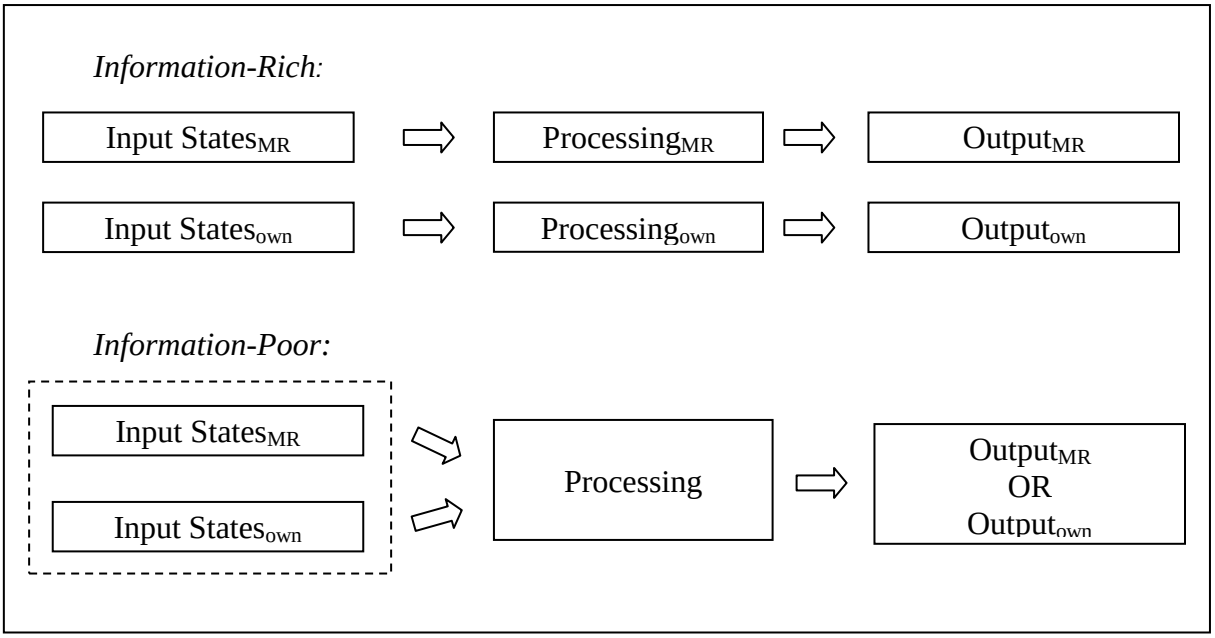


[Figure 1: Functional Sketch of the Mindreading System]

It is important to note that there is no a priori reason to assume that the two parts of this mechanism (i.e. the generator and the processor) have to have the same 'informational' structure. In fact, much recent work has moved away from analysing the mindreading system as either *completely* information-rich or *completely* information-poor, and has instead come to see it as being composed of different parts with different informational natures (see e.g. Goldman 2006; Nichols and Stich 2003; Carruthers 2006, 174-176). This will become important again below.

[3] Note that the mindreading procedure sketched in figure 1 might have to be performed more than once in any one application of the system (see also Gordon 1986; Goldman 2006, 183-185). Note also that – to avoid terminological complications – the processor is to be seen as doing more than merely *repeating* the initially generated mental states.

Given this (admittedly quite crude) sketch, it now becomes possible to make more precise the difference between the two views of the mindreading system. This difference can be formulated as follows: is it the case that the mindreading system is layered *on top of* what is in the agent's cognitive architecture anyway, or is it the case that it (at least partially) overlaps with systems that that are also active in the agent's *own* cognitive life? Graphically, this difference can be expressed as follows [figure 2]:



[Figure 2: Two Competing Mindreading Architectures]

What this figure makes clear is that, on the information-rich picture, what happens when an agent predicts a decision or an inference and when she makes a decision or an inference herself is entirely different: the two situations make use of completely distinct systems. By contrast, on the information-poor picture, some or all of the components of the mindreading system are *shared* with the relevant components of the agent's own cognitive life: in particular, while the

two generators may or may not be dedicated to their respective tasks (see the dotted line in figure 2) – an issue that is still open for debate – it is at least the case that there is only *one* processor (see also Goldman 2006, 26-29).

Two further points concerning this way of setting up the issue are worth making here. Firstly, while I here maintain the common labels associated with the two accounts of mindreading, it is not at all clear that these labels are really useful for signposting the distinction between these accounts (see also Nichols and Stich 2003, 132-135). In particular, it is not at all obvious that the fundamental issue to be debated concerns how much *information* there is in the mindreading system, or whether a *simulation* or a *theory* is used in predicting decisions or inferences.[4] Because of this, these labels should here be taken to be merely *proper names* – and not *descriptions* – of the two competing theories.

Secondly, the focus here is squarely on what has become known as 'high-level' mindreading (see e.g. Goldman 2006, chap. 7): the attribution of complex and cognitively demanding mental states like the propositional attitudes. This contrasts with 'low-level' mindreading (see e.g. Goldman 2006, chap. 6): the attribution of emotions, action-intentions, and other phylogenetically, neurally, and cognitively more basic mental states. There are two reasons for this focus on high-level mindreading: on the one hand, it makes the discussion more compact, and on the other, it has much inherent interest – some of the most problematic issues surrounding the nature of the mindreading system concern the high-level variety (see e.g. Goldman 2006). Accordingly, unless explicitly noted, appeals to 'mindreading' here should always be understood as referring to *high-level* mindreading only.

[4] This becomes clear e.g. from noting that this would require either a cogent distinction between 'explicitly' and 'implicitly' represented information, or a suitably well-motivated definition of 'simulation' – neither of which has turned out to be easy to find (see e.g. Clark 1992; Goldman 2006).

With the dispute between information-rich and information-poor accounts of mindreading thus clarified, it is now possible to turn to evolutionary theory to see what it suggests concerning it. This is the topic of the next section.

3. Natural Selection and the Argument from Simplicity

For reasons that will be defended in more detail in section 4, it is sufficient here to *assume* that

(i) mindreading evolved (primarily) by natural selection

(ii) mindreading was adaptive for helping us deal more successfully with other organisms than we could do beforehand.

While (i) and (ii) certainly deserve much further discussion—after all, a trait can evolve in many different ways, with (unconstrained) natural selection merely being one option among many—an argument based on them can still be very telling. Primarily, this is because, even in the absence of certainty about the truth of (i) and (ii), determining what they entail about the structure of the mindreading mechanism can have important implications for the further study of the latter. In particular, as will also be made clearer below, finding out about the consequences of (i) and (ii) can give us an important *novel handle* on the investigation of our mindreading system.

Furthermore, for present purposes, it is useful to restrict this investigation of the evolution of the mindreading mechanism to the mindreading *processor*. This is helpful, as it makes the discussion to follow more tractable: the issues here are complex, and considering the processor on its own allows them to be treated in more depth and with more clarity than would otherwise be possible. Of course, this is not to deny that the processor and the generator are interconnected,

and that successful mindreading depends on both of them working together. All it means is that I seek to assess the selective pressures on the processor separately from those on the generator – which, given the gradual nature of evolution and the overall structure of the mindreading system, should not seem to be greatly implausible.[5]

What, then, do (i) and (ii) above imply about the informational nature of the mindreading processor? A frequently hinted-at answer to this question is that the processor is information-poor, as this makes the mindreading system more 'parsimonious' (see also Harris 1992):

> "A fundamental idea of ST [Simulation Theory] is that mindreaders capitalise on the fact that they themselves are decision makers, hence possessors of decision-making capacities. To read the minds of others, they need not consult a special chapter on human psychology, containing a theory about the human decision-making mechanism. Because they have one of those mechanisms themselves, they can simply *run* their mechanism on the pretend input appropriate to the target's initial position."
> (Goldman 2006, 20; italics in the original)

> "This is one of the places where our account borrows an idea from simulationist theorists […] Why would Mother Nature go to the effort of arranging for us to have a *theory* about how the inference mechanism works when she could get the job done by using the inference mechanism itself?"
> (Nichols and Stich 2003, 104; italics in the original).

I think the basic idea behind these quotes is fundamentally correct; however, as it stands, it is neither precise enough nor complete enough to make a compelling case for information-poor mindreading. In what follows, I shall try to spell it out in more detail.

To do this, it is necessary to begin by making clearer what the starting point of the evolution of mindreading was: it is only by knowing what mental feats our ancestors were able to perform before they started mindreading that we can come to understand the selective pressures that led them to acquire the latter ability.

[5] This also gains strength from the recent popularity of *hybrid* theories of mindreading (as mentioned earlier). This popularity corroborates the fact that the two components of the mindreading system are not so intertwined that a separate discussion of them is impossible.

Now, the most widely accepted account of the structure of our pre-mindreading minds is that they are (at least in a rough-and-ready way) well described by our common-sense *belief / desire psychology* (see e.g. Papineau 2001; Nichols and Stich 2003, 13, 60-66; Carruthers 2006, 65-68; Goldman 2006). The major reason for thinking this is that having a representationalist mind – comprising conative states, doxastic states, and various mechanisms for combining them – is adaptive in a wide range of circumstances, and that for reasons having nothing to do with mindreading at all (see e.g. Godfrey-Smith 1996; Sterelny 2003; see also Sober 1994; Stephens 2001; Smead and Zollman forthcoming).[6]

Given this, it becomes possible to formulate a more precise version of the above argument for the information-poor theory of mindreading. One way to do so is as follows.[7]

(*The Argument from Simplicity*)

1. An agent's decision and inference making system can also be used (by her) for the prediction of another agent's decisions and inferences.
2. Building extra cognitive systems is costly, *ceteris paribus*.
3. Therefore, natural selection favours the avoidance of a separate, dedicated system for decision and inference prediction, *ceteris paribus*.

[6] An objection that is sometimes raised to this argument claims that the above is the *reverse* of the true chronology: that is, it is argued that the evolution of our belief / desire psychology was actually driven by the evolution of our mindreading abilities (see e.g. Godfrey-Smith 2003; Sterelny 2003, chap. 4). However, for two reasons, this objection is not very damaging to the present discussion. On the one hand, while it might be plausible that our minds received considerable *sophistication* due to our interactions with other agents, it is not clear that the basic components of our mind could have evolved *entirely* in this way (see also Godfrey-Smith 2003). On the other hand, even if this were a coherent possibility, the rest of the discussion would not be impacted all that much: instead of needing to ask if natural selection favours a dedicated *mindreading system*, we would then need to ask if it favours a dedicated *inference and decision making system*. This is bound to bring up many of the same issues that are raised in what follows.

[7] Note that the form of the following argument is *contrastive*: the claim is not that building a dedicated mindreading processor is unlikely *per se*, but only that it is less likely than the competing, more parsimonious alternative (see also Sober and Wilson 1998). While it is true that this more parsimonious alternative is not *costless* (for example, a way has to be found to feed the beliefs and desires attributed to the target into the agent's own reasoning mechanisms), it is still bound to be vastly cheaper *overall* than duplicating an entire practical and theoretical reasoning system.

In more detail, this argument can be laid out as follows.[8]

The first premiss takes off from the fact that the agent's own reasoning system and her mindreading processor need to solve what are essentially identical problems – namely, deriving appropriate decisions and inferences (either actual ones or merely predicted ones) from a given set of beliefs and desires (either the agent's own or those attributed to the target). Because of this, it seems highly plausible that the agent's own decision and inference making system can also play the role of the mindreading processor: the capabilities needed to fulfil the functions of either system are largely isomorphic. For this reason, premiss 1 of the argument from simplicity should be seen to be quite uncontroversial.

The second premiss is an expression of the (equally uncontroversial) fact that it is costly for an organism to assemble and maintain extra cognitive systems. There are three main reasons for this. Firstly, building and maintaining an extra cognitive system requires building and maintaining a novel store of knowledge; in turn, this increases the organism's memory demands and slows down computations. Secondly, an extra cognitive system requires extra energy to be usable. Thirdly, to be functional, any extra system needs to be integrated into the rest of the cognitive architecture, thus requiring (potentially very costly) alterations to some of the latter's existing elements (see also Carruthers 2006, 74).

From these premises, it follows straightforwardly that natural selection will avoid the duplication of the agent's own reasoning systems for mindreading purposes, *unless the costs of this duplication are outweighed by further adaptive benefits*. Since the agent's own reasoning systems can be used for mindreading processing (by premiss 1), and since natural selection opts for more efficient design solutions over less efficient ones (by premiss 2), there is reason to think that natural selection will favour an information-poor mindreading system, *ceteris paribus*.

[8] I thank an anonymous referee for some useful remarks about this argument and the discussion concerning it.

It is with this conclusion, however, where the argument runs into difficulties: the ceteris paribus qualification inherent in it makes it necessary to ensure that things are, in fact, the same. This is especially important here, as there is an antecedent reason to think that things may *not* be the same: namely, there are many cases where the duplication of existing cognitive systems yields *benefits* (see also Nichols and Stich 2003, 105). If these benefits are great enough to outweigh the costs of the duplication, natural selection should therefore *favour* duplication, not *disfavour* it. In turn, this means that, in order for the argument from simplicity to be plausible, it becomes necessary to make sure that the duplication of the agent's own reasoning systems does *not*, in fact, lead to these kinds of benefits.

These benefits fall into two main categories: on the one hand, duplication is useful where it allows for *parallel processing*, as that makes for much added computational speed (see e.g. Carruthers 2006, 24-25). On the other, duplication is useful where it allows for the construction of *back-up* systems for important abilities, as this insures the organism against catastrophic drops in fitness (see e.g. Sober and Wilson 1998, 320; Stich 2007; Schulz forthcoming). Consider these two potential benefits in turn.

1. Parallel Processing

The first potential benefit of an information-rich design for the mindreading system is that it would allow for the simultaneous processing of mindreading problems and the agent's own decision and inference problems. In situations where time is pressing (and it seems reasonable to think that these situations are quite widespread in the natural world), such parallel processing may be highly adaptive: it would allow an organism to simultaneously *decide* what to do in a certain situation and to *predict* how others are going to react to her decision.

However, for two reasons, it seems implausible to think that the possibility of this kind of parallel processing led natural selection to favour information-rich mindreading. Firstly, there are empirical considerations suggesting that, *in fact*, this was not the case; secondly, there are theoretical considerations suggesting that this is not something natural selection should be *expected* to favour here. Begin with the empirical side.

There are good reasons to think that we as *a matter of fact* struggle to mindread and make decisions at the same time. While there is no empirical work directly pertaining to this issue, there are a number of studies that are at least indirectly concerned with it: in particular, there is evidence suggesting that the mindreading system and the agent's own decision / inference making system can be made to interfere with each other (see e.g. Bull et al. 2008; see also Apperly et al. 2008; Leslie 2000).

This matters, since it is highly plausible that natural selection actually *was* a key factor influencing the evolution of our mindreading abilities (as will also be made clearer in section 4). For this reason, this kind of finding makes for an indirect argument against the claim that natural selection favoured the parallel processing of mindreading and decision / inference problems: for if natural selection did favour this, one would expect to find evidence for it. Since we in fact find evidence for its *absence*, however, it is more plausible to think that this is *not* what natural selection favoured. Of course, since it has not yet been *conclusively* established that natural selection was the only factor influencing the evolution of high-level mindreading, this sort of argument can only be seen to give tentative support to the argument from simplicity. However, it is possible to strengthen this argument further by noting that there is also a theoretical reason for thinking that natural selection did not favour parallel processing in this case.

To see this, note that, by assumption (ii) above, the major benefit of accurate mindreading comes from improved social interactions – i.e. from improvements in our *behaviour* vis-à-vis other organisms. In turn, this entails that mindreading must go beyond pure 'behaviour-reading' (see e.g. Dennett 1978, 275): it must do more than merely predict and explain a target's behaviour using various *empirical generalisations*, but must make reference to the *mental causes* of this behaviour. This is due to the fact that pure behaviour-reading can also be achieved by using a belief / desire psychology that lacks mindreading abilities: after all, this kind of 'social cognition' does not differ fundamentally from the prediction of the movements of inanimate objects or simple animals.

In turn, this suggests that for accurate mindreading to be adaptive, it is necessary that we incorporate the *predicted* decisions and inferences of the target into our *own* decisions and inferences. In other words, for mindreading abilities to be selected for due to their helping us to improve making behavioural predictions, mindreading and decision / inference problems need to be processed *serially*: we first have to *collect* the data about the target before we can *use* them in our own decision and inference making mechanisms. For this reason, parallel processing does not appear to be actually *wanted* when it comes to decision / inference making and mindreading: the former is meant to *contain* the latter as one of its elements; accordingly, the latter should be completed before the former is begun. To fully appreciate this point, two objections to it ought to be considered here.

Firstly, it may be thought that it is possible to compute, in parallel, *provisional* answers to the decision / inference problem and the mindreading problem, and then to combine these answers afterwards to arrive at the overall solution sought for. In this way, a somewhat weakened form of parallel processing may seem to be salvageable in this context: while the mindreading and the

agent's own reasoning problems cannot *strictly* be solved in parallel, they can at least be solved in a way that contains *parallel elements*.[9]

However, for two reasons, this objection is unlikely to completely invalidate the conclusion derived earlier. On the one hand, it is not yet fully clear that the above 'provisional' form of parallel processing can actually be made to work: excluding an important element of the decision / inference making process – even if only provisionally – can completely alter the outcome of the latter, and that in a way that cannot easily be rectified in retrospection. Primarily, this is because reasons for action and thinking are not generally 'additive': if A is the best action relative to considerations R_1 to R_n, it need not also be the best action relative to considerations R_1 to R_{n+1}.

On the other hand, even if this sort of parallel processing does turn out to be feasible, it is not yet clear that it would yield great processing advantages. Given the complexity of our social environment, the necessary revisions to our provisional decisions or inferences are likely to be extensive; in turn, this suggests that this kind of parallel processing will not be much faster than serial processing would be – thereby doing away with the former's major benefits. At the very least, therefore, more work is needed to make this a fully compelling challenge to the above argument.

The second objection to be considered here argues that it may be adaptive to be able to mindread someone else for 'future use', while making a different decision *now*. In this way, it may again seem that a form of parallel processing may be justifiable in this context: while it might be granted that a mindreading problem cannot be solved in parallel with the decision / inference problem *to which it is directly relevant*, it might also be argued that it can still be solved in parallel with a *different* decision / inference problem.

[9] I thank an anonymous referee for some useful remarks concerning this issue.

However, this objection is also unlikely to completely invalidate the above conclusion. The reason for this is that it is not clear that individual decision / inference problems can always be so easily separated from one other.[10] In particular, instead of seeing an agent as facing a succession of individual decision problems, it will frequently be more appropriate to see her as facing a single decision problem about an entire *course of action* (i.e. about what she is to do from now on into the future). Importantly, this will be particularly plausible in the case of an agent that seeks to mindread someone else for future use: after all, the reason why she is mindreading in this way is most likely precisely the fact that she is preparing for a particular future action.[11] This, though, means that the present objection has not actually addressed the above argument – which is that, *given the appropriate time horizon*, it is not feasible to solve decision problems and mindreading problems at the same time.

Overall, therefore, it becomes clear that there is little reason to think that the benefits of parallel processing dwarf the disadvantages of the duplication of the agent's own decision and inference making system. Consider, then, the possibility that this duplication was instead beneficial for leading to the existence of a backup for an important cognitive system.

2. Backup Systems

Is it the case that natural selection favoured the duplication of the agent's own reasoning systems for mindreading purposes, since the two systems are useful backups for each other? That is, did it favour information-rich mindreading, as this allowed the mindreading processor to be used for decision and inference-making purposes when the agent's primary systems for this were

[10] This is a well-known problem in Rational Choice Theory – see e.g. Savage (1954, 25-26; 82-91).
[11] This becomes especially obvious when it is noted that there is an indefinitely large set of possible future decision and inference problems. Unless the agent has some idea about which of these are actually relevant to her, indiscriminate 'preparatory mindreading' would completely swamp her cognitive system.

damaged (and vice versa)? This possibility must be taken seriously, as having a backup for an important system can be very useful, and is something that natural selection has often opted for (see also Sober and Wilson 1998; but also Stich 2007 and Schulz forthcoming). However, there are three reasons to think that it did not do so in the present context.

Firstly, it seems to be true that, as a matter of fact, the two systems are *not* used as backups for one another. As in the case of parallel processing, this is important, since it is very likely that natural selection played a major role in the evolution of high-level mindreading. Given this, we again would expect to find evidence of the kind of design it favoured – i.e. we would expect to find evidence for the mindreading system being usable as a backup for the agent's own reasoning systems. As it turns out, however, this is not the case.

In general, the main type of empirical finding that would seem to bear on this issue concerns psychopathologies (autism, Down Syndrome, Williams Syndrome, and schizophrenia). Unfortunately, this evidence is very hard to make sense of: on the one hand, there is hardly any empirical work concerning these pathologies whose interpretation is obvious or widely agreed on (for example, the hypothesis that autism can be seen as 'mindblindness' is fairly controversial and ought not to be taken as fully corroborated yet – see e.g. Baron-Cohen 2002; Klin et al. 2008; see also Goldman 2006, 200-205). On the other, most of this evidence is anyway best analysed as concerning the failure of the mindreading *generator* – i.e. the component of the mindreading system that is not at issue here (this is true e.g. of the work of Leslie 2000 and Baron-Cohen 2002).

From an empirical standpoint, therefore, there is little reason to take seriously the view that natural selection sought to make the mindreading processor and the agent's own decision /

inference making systems usable as backups for each other. This conclusion is further strengthened by the fact that several other considerations support it as well.

On the one hand, there may well be *design constraints* on the cognitive architecture in question that make this kind of 'emergency coupling' difficult or impossible. That is, it may well be difficult or impossible to construct the two systems in such a way that they take different inputs (beliefs and desires attributed to the target as opposed to the agent's own beliefs and desires), yield different outputs (predicted decisions as opposed to the agent's own decisions), and can still be substituted for each other if necessary. At the very least, much more needs to be said about *how* such an architecture is meant to work – as things stand, it is just not clear that this was even an option that natural selection *could have* favoured.

On the other hand, there are also reasons to think that natural selection favours the neural co-location of the mindreading processor and the agent's own reasoning systems.[12] Leaving aside for a moment what these reasons are, it needs to be noted that the existence of these reasons means that it is very likely that the two systems will break down in exactly the same circumstances. In turn, this makes them bad backups, as it will then be very likely that the secondary system is unusable when the primary system has broken down – i.e. exactly when it is most needed. In this way, it becomes clear that any reason for thinking that natural selection opted for neural overlap between the mindreading processor and the agent's own reasoning systems will at the same time be a reason for thinking that it favoured them *not* to be used as backup systems for each other. As it happens, there are two reasons of precisely this form.[13]

[12] It is important to keep in mind here and in what follows that *neural* overlap is not the same as *functional* overlap: whether the latter is true for the mindreading processor and the agent's own reasoning systems is a different question from whether the former is true. See also note 1.

[13] Saxe et al. (2006) and Saxe & Powell (2006) might seem to suggest that there is no such neural overlap. However, when considering their studies more closely, it becomes clear that they do *not* show this: their findings are really only concerned with the mindreading *generator*, not the processor, and hence do not bear on the present issue.

Firstly, neural overlap is made plausible by the fact that our decision and inference making systems are *cognitively and neurally complex*. These systems are at the heart of what our brain is for – they comprise the two key features of much of human thought (see e.g. Fodor 1983; Carruthers 2006). Because of this, it is quite likely that they require large amounts of neural resources to function. Since the overall quantity of these resources is limited, it would thus not be surprising if the two systems *shared* many of them.

Secondly and more importantly, neural overlap is exactly what we find when it comes to an analogous system: the *low-level* mindreading system. This matters, as this system really did seem to have evolved mostly by natural selection (see e.g. Goldman 2006, chap. 10; de Waal 2008); in turn, this suggests that natural selection does seem to have favoured the neural co-location of the systems responsible for the *attribution* of low-level mental states and for the generation of the agent's *own* low-level mental states (see e.g. Goldman 2006, chap. 5; Wicker et al. 2003).[14] Further, it is hard to see why natural selection would not operate in the same way when it comes to high-level mindreading, as both of these systems have to deal with the same issues: in particular, it would surely be useful – at least prima facie – to have a second set of 'emotional centres' as a backup, in case the primary ones break down. However, natural selection seems to have legislated against this, and opted for neural overlap instead. While not *entailing* that it would do the same when it comes to high-level mindreading, this does seem to give at least *some reason* to think so.

In sum: it is also fairly unconvincing to think that natural selection might have favoured information-rich mindreading due to the benefits of having backup-systems. This is empirically implausible, may not be a design option, and conflicts with the fact that natural selection seems

[14] This is accepted even by Saxe (forthcoming).

to have favoured neural overlap between the mindreading processor and the agent's own decision and inference making systems.

On the whole, therefore, the natural selection-based perspective developed here supports the argument from simplicity: decision and inference prediction ought to be based on the agent's own practical and theoretical reasoning systems, as this avoids the latter's costly duplication. Before concluding, it is useful to briefly make clearer what role these sorts of evolutionary considerations can play in the discussion surrounding high-level mindreading.

4. The Power of Arguments from Natural Selection

What is the value of establishing what natural selection has to say about our mindreading abilities? As noted earlier, it cannot simply be assumed that natural selection truly was the major influence on the evolution of these abilities – especially in small, finite populations (which were characteristic of the early hominid lineage), other evolutionary factors such as drift and various genetic constraints need to be taken into account as well. Moreover, even if natural selection really was of prime importance here, it must be noted that this factor is still constrained by our evolutionary history, and cannot simply start from scratch in designing our minds. For these reasons, it might seem that, until it has been established that (a) natural selection was, in fact, the most important factor driving the evolution of high-level mindreading, and (b) there were no major historical constraints affecting its workings, arguments based on what natural selection *would favour* do not have much epistemic worth.

However, things are not quite as negative as that. There are in fact numerous reasons for taking natural selection-based arguments like the one above seriously, even in the absence of a full corroboration of (i) and (ii). Three of these are particularly important here.

Firstly, when it comes to high-level mindreading, the emphasis on natural selection is a highly compelling (and widely accepted) evolutionary hypothesis: given the complexity of our mindreading abilities, adaptationist explanations of their evolution should loom large (see e.g. Cosmides and Tooby 1992; Pinker and Bloom 1990; Dawkins 1986). Furthermore, the particular adaptive scenario chosen here – i.e. the one that focuses on improved social interactions – is among the most plausible ones available, and also widely endorsed in the literature (see e.g. Byrne and Whiten 1988; Sterelny 2003). Lastly, as yet, we are not aware of any specific historical preconditions that might have fundamentally altered the working of natural selection in this case. In short: even though (i) and (ii) cannot yet be considered fully corroborated, they do represent highly plausible hypotheses.

Secondly, arguments from natural selection like the one above can have evidential significance even in the absence of a full confirmation of the equivalents of (i) and (ii). While these arguments are not strong enough to force us to *accept* one theory or another, they are still strong enough to present *evidence* in favour of one theory over another – i.e. they can point to factors that tilt the balance of reasons towards one particular side (see also Sober and Wilson 1998). In particular, in the present context, this kind of argument can make clear that there are interesting and important considerations speaking in favour of information-poor mindreading. While it is not yet fully clear *how* important these considerations are, their existence must still be admitted to be relevant to the debate surrounding high-level mindreading – especially given the fact that other sources of evidence concerning this matter are quite ambiguous.

Finally, the argument given above is also useful for purely methodological (i.e. non-evidential) reasons: it can help make clearer what studies and investigations ought to be done in order to get more evidence about the nature of the mindreading mechanism. For example, the above makes clear that it would be beneficial for the resolution of the debate between information-rich and information-poor theories of mindreading to have (more) evidence concerning:

1. our ability to solve mindreading and decision / inference making problems in parallel.
2. the general possibility of embedding two systems in our minds in such a way that they are usable as backups, but that their inputs can still be kept strictly apart in the course of their normal operation.
3. the neural localisation of the high-level mindreading and the decision / inference making systems.

In this context, it is also important to note that it is in particular the above argument that brings out the importance of these tests. This does not mean that these tests could not also be derived in other ways (especially 1 and 3); the point is just that, without the above argument, it might have been much harder to see, firstly, *that* they are relevant here, and secondly, *how* they are relevant here.

5. Conclusion

I hope to have shown that, assuming high-level mindreading was the target of natural selection due to its leading to improved social interactions, there are good reasons for thinking that it is (at

least partly) information-poor in nature. Specifically, I hope to have shown that the common argument that natural selection disfavours the mindreading processor from being distinct from the agent's own reasoning systems is plausible – but also that matters here are much more complicated than is frequently supposed. In particular, I hope to have shown that it is also necessary to consider whether the costs of a potential duplication of this processor are balanced by potential benefits from parallel processing or the existence of backup systems. While this does not seem to have been the case, it remains true that it is only by taking these issues into account that the argument from simplicity can be made plausible at all. Finally, in this way, I also hope to have shown how evolutionary considerations can inform philosophical and psychological theorising more directly than by merely providing 'just so stories'. Overall, therefore, while certainly not ending the dispute concerning the nature of high-level mindreading, the evolutionary argument presented here does present a step towards its resolution.

Acknowledgements

I would like to thank Alvin Goldman, Elliott Sober, Larry Shapiro, Stephen Stich, Eric Margolis, and an anonymous referee of this journal for useful comments on previous versions of this paper.

References

Apperly, I., Back, E., Samson, D., & France, L. (2008). The Cost of Thinking about False Beliefs: Evidence from Adults' Performance on a Non-Inferential Theory of Mind Task. Cognition, 106, 1093-1108

Baron-Cohen, S. (2002). The Extreme Male-Brain Theory of Autism. Trends in Cognitive Science, 6, 248-254.

Byrne, S., & Whiten, A. (1988) (eds.). Machiavellian Intelligence. Oxford: Oxford University Press.

Bull, R., Phillips, L., & Conway, C. (2008). The Role of Control Functions in Mentalizing: Dual-Task Studies of Theory of Mind and Executive Function. Cognition, 107, 663-672.

Carruthers, P. (1996). Language, Thought and Consciousness. Cambridge: Cambridge University Press.

Carruthers, P. (2006). The Architecture of the Mind. Oxford: Oxford University Press.

Clark, A. (1992). The Presence of a Symbol. Connection Science, 4, 193-205.

Cosmides, L., & Tooby, J. (1992). The Psychological Foundations of Culture. In J. Barkow, L. Cosmides and J. Tooby (eds.), The Adapted Mind (pp. 19-136). Oxford: Oxford University Press.

Dawkins, R. (1986). The Blind Watchmaker. Oxford: Oxford University Press.

Dennett, D. (1978). Brainstorms. Montgomery: Bradford.

de Waal, F. (2008). Putting the Altruism back into Altruism: the Evolution of Empathy. Annual Review of Psychology, 59, 279-300.

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

Fodor, J. (1992). A Theory of the Child's Theory of Mind. Cognition 44, 283-96.

Godfrey-Smith, P. (1996). Complexity and the Function of Mind in Nature. Cambridge: Cambridge University Press.

Godfrey-Smith, P. (2003). Untangling the Evolution of Mental Representation. In A. Zilhao (ed.), Cognition, Evolution, and Rationality: A Cognitive Science for the 21st Century. London: Routledge.

Goldman, A. (2006). Simulating Minds. Oxford: Oxford University Press.

Gopnik, A., & Wellman, H. (1994). The Theory Theory. In L. Hirschfeld & S. Gelman (eds.), Mapping the Mind (pp. 257-293). Cambridge: Cambridge University Press.

Gordon, R. (1986). Folk Psychology as Simulation. Mind & Language, 1, 158-172.

Harris, P. (1992). From Simulation to Folk Psychology: The Case for Development. Mind and Language, 7, 120-144.

Klin, A., Volkmar, F., & Sparrow, S. (2008). Autistic Social Dysfunction: Some Limitations of the Theory of Mind Hypothesis. Journal of Child Psychology and Psychiatry, 33, 861 – 876.

Leslie, A. (2000). 'Theory of Mind' as a Mechanism of Selective Attention. In M. Gazzaniga (ed.), The New Cognitive Neurosciences (pp. 1235-1247). 2nd Edition. Cambridge, MA: MIT Press.

Nichols, S., & and Stich, S. (2003). Mindreading. Oxford: Oxford University Press.

Papineau, D. (2001). The Evolution of Means-End Reasoning. In D. Walsh (ed.), Naturalism, Evolution, and Mind (pp. 145-178). Cambridge: Cambridge University Press.

Savage, L. (1954). Foundations of Statistics. New York: John Wiley.

Saxe, R. (forthcoming). The Neural Evidence for Simulation is Weaker than I Think You Think It Is. Philosophical Studies.

Saxe, R., and Powell, L. (2006). It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. Psychological Science, 17, 692-698.

Saxe, R., Schulz, L., & Jiang, Y. (2006). Reading Minds versus Following Rules: Dissociating Theory of Mind and Executive Control in the Brain. Social Neuroscience, 1, 284-98.

Schulz, Armin (forthcoming). Sober & Wilson's Evolutionary Arguments for Psychological Altruism: A Reassessment. Biology and Philosophy.

Smead, R., & Zollman, K. (forthcoming). Language and the Baldwin Effect. Philosophical Studies.

Sober, E. (1994). The Adaptive Advantage of Learning and A Priori Prejudice. In From A Biological Point of View (pp. 50-70). Cambridge: Cambridge University Press.

Sober, E., & Wilson, D. S. (1998). Unto Others: The Evolution and Psychology of Unselfish Behavior. Cambridge, MA: Harvard University Press.

Stephens, C. (2001). When Is It Selectively Advantageous to Have True Beliefs?. Philosophical Studies, 105, 161-189.

Sterelny, K. (2003). Thought in a Hostile World: The Evolution of Human Cognition. Oxford: Blackwell Publishing.

Stich, S. (2007). Evolution, Altruism and Cognitive Architecture: A Critique of Sober and Wilson's Argument for Psychological Altruism. Biology and Philosophy, 22, 267-281.

Stich, S. (2009). Response to Egan. In D. Murphy & M. Bishop (ed.), Stich and his Critics. New York: Wiley.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P, Gallese, V., & Rizzolatti, G. (2003). Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust. Neuron, 40, 655–664.