Reply to Carruthers, P., Fletcher, L. & Ritchie, B. "Evolving Self-Consciousness"

1. Summary

I learned a good number of things from reading this paper, but I remain puzzled about a good number more. Carruthers et. al. contrast two accounts of the evolutionary function of self-consciousness. According to one, the Control Account (CA), self-consciousness is an adaptation for the monitoring and control of the subject's own psychological states. As they put it, "self-consciousness evolved for the purposes of metacognitive monitoring and control" (p.2). According to the other, the Mindreading Account (MRA), self-consciousness is not an adaptation at all, but rather an exaptation. It is simply the turning towards ourselves of the mindreading faculty that is an adaptation to independent, social pressures. As they put it, "the adaptation underlying the capacity for knowledge of one's own mental states is the mindreading faculty […] which evolved initially for social purposes." (p.3) Carruthers et. al. argue that CA predicts a 'native competence' for monitoring and control, emerging in the second year of life, and buffered against individual variation (pp.7-9). This prediction, they argue, is not confirmed (pp.9-12). MRA, it is argued, also makes a number of predictions: that humans possess a native competence for mindreading, emerging in the second year, and buffered against individual variation; that metacognitive abilities cannot outstrip mindreading abilities; that at any given stage of development, subjects will exhibit an equal level of competence in self-conscious and mindreading abilities; and that whilst subjects will be good at monitoring their psychological states, there is no native competence for the control of such states (pp.5-7). These predictions, argue Carruthers et. al. are confirmed (pp.9-17). The conclusion is that "the latter account (MRA) is currently the best supported of the two" (p.1).[1]

2. What is self-consciousness?

Carruthers et. al. seem to use the terms 'self-consciousness', 'self-awareness', 'self-knowledge', and 'metacognition' interchangeably. Indeed, the paper opens by introducing three of these notions, apparently without any significant differences of meaning, "Humans have the capacity for

[1] There is a question over exactly what Carruthers et. al. take their conclusion to be. On the face of it, the conclusion is that MRA has stronger empirical support than CA. However, they also say that all existing first-person based views are undermined by their arguments against the plausibility of CA (p.3). If by this they mean all views of the evolutionary function of self-consciousness that accord a priority to first-personal access, then this certainly doesn't follow from the arguments presented, since not all such views will share the objectionable features of CA (the predictions concerning control). It is, then, better to interpret them as referring to all variations of CA. These might include, for example, the view that self-consciousness is an adaptation for monitoring and control and which is *also* subsequently employed in mindreading abilities. So interpreted, whilst the claim seems perfectly reasonable, it is worth pointing out that it doesn't undercut views such as Goldman's (2006) simulationism, since he doesn't (as far as I am aware) endorse the control hypothesis distinctive of CA.

awareness of many aspects of their own mental lives—their own experiences, feelings, judgments, desires and decisions. We can often ***know*** what it is that we see, hear, feel, judge, want or decide. This article examines the evolutionary origins of this form of ***self-consciousness***." (p.1, my emphasis). They also seem to treat self-consciousness as involving the *conceptual* representation of oneself. This is suggested by the identification of self-consciousness with self-knowledge, but also with the claim that it involves awareness *that* we are in some psychological state, "We don't just see, we are aware *that* we see" (p.1). Given this, one would expect self-consciousness to be manifest in the capacity to be in states with contents of the form 'I am Ψ'. There is, finally, another reason for thinking that self-consciousness, as conceived by Carruthers et. al., involves conceptual representations of this sort. In their characterisation of MRA, Carruthers et. al. describe the way in which the mindreading system can play a metacognitive role as follows, "[t]he mindreading faculty would have access to globally broadcast perceptual and imagistic representations as input, and attributions of mental states to oneself would initially utilize this input together with the same core knowledge and principles that are employed for third-person mindreading." (p.3). How do they conceive of the 'access' that the mindreading capacity has to these internal states? Surely *not* as an exercise of self-consciousness for, if it were, then the mooted explanation of self-consciousness would itself rely on an unexplained capacity for self-consciousness. Yet one might well suppose that the access in question would involve a (perhaps non-conscious) representation of the subject as being in some psychological state. How might this be distinguished from self-consciousness proper? The most obvious way to distinguish this capacity from self-consciousness would be to treat it as a non-conceptual form of representation, reserving the term 'self-consciousness' for a conceptual representation of oneself as being in some psychological state, something of the form 'I am Ψ'.[2]

3. What predicts what?

The descriptions of CA and MRA offered by Carruthers et. al. are very minimal. As a result, one may suspect that, as they stand, they might not really make all the predictions identified. With respect to CA, however, the minimal nature of the account is matched by that of the predictions and, although the timing prediction does seem to me to be somewhat speculative, it is not crucial to Carruthers et.

[2] To this it may be objected that the access in question need not involve representations of psychological states *as* my own but can get by simply with representations of psychological states *of* my own. Perhaps this is related to Carruthers' claim that, "Receiving as input a visual representation of a person smiling, for example, the mindreading system should be capable for forming the judgement, I AM SEEING A PERSON SMILING […] there is simply no need to postulate any separate faculty of 'inner sense.'" (Carruthers 2011, p.51) But if this move is made, it strikes me that we are at the very least owed an explanation of how it is that the mindreading faculty's representation of a psychological state that happens to be mine can produce the judgement that the state belongs to *me*.

al.'s case, so I will accept it and the others for the sake of argument. Nor will I challenge the claim that these predictions are disconfirmed.

Things are less clear cut with respect to MRA. We can break down the view into two parts: (MR) The mindreading faculty is an adaptation for various social purposes, and (MR-SC) Self-consciousness simply involves the mindreading faculty turned to oneself. Now, some of the predictions identified by Carruthers et. al. result from MR and some from MR-SC. The prediction that we have a native competence for mindreading, emerging in the second year, and buffered against individual variation results from MR. The predictions that self-conscious abilities cannot outstrip mindreading abilities, that they will exhibit equal competence, and that control of psychological states is not a native competence, all result from MR-SC.

Furthermore, MR is consistent with CA, which is silent about the evolution of the mindreading faculty (p.3). What is really distinctive, then, about MRA is the MR-SC component, the claim that self-consciousness is nothing more than self-directed mindreading. As a result, it is not obvious why we should think that evidence for MR, in the absence for evidence for MR-SC, counts as support for MRA. Evidence for MR is not evidence for MR-SC, or even evidence for the conjunction of MR and MR-SC. Evidence for MR is no more evidence for MRA than it is for the conjunction of MR and CA. Of course, it may be that MRA is a simpler, more economical theory than the conjunction of MR and CA, and perhaps that counts in its favour. But this doesn't seem to be an argument that Carruthers et. al. make. If this is right, then a significant amount of the case mounted for MRA, doesn't really distinctively support that view at all.

What, then, of the predictions that result directly from MR-SC? Recall that there are three: *No outstripping* is the claim that, "there are no creatures capable of self-consciousness of a sort that cannot be explained in terms of whatever degree of mindreading ability those creatures also possess" (p.17)[3]; *Equal competence* is the claim that, "[i]f people are successful mindreaders […]

[3] At one point, Carruthers et. al. formulate this constraint rather differently, "infants will be capable of self-consciousness as soon as they become capable of third-person mindreading" (p.6). The idea is that if self-consciousness is just mindreading that is directed towards oneself, then what one can do for others, one will be able to do for oneself. If, as I suggested above, the exercise of self-consciousness involves the employment of the first-person concept, then this claim relies on a substantial assumption. The outputs of the mindreading faculty in third-person cases will be attributions of psychological states to others, for example, "She is Ψ". The outputs of the mindreading faculty in the first-person case, on the other hand, will be attributions of psychological states to oneself, representations with content of the form, "I am Ψ". The latter employ the first-person concept, the former do not. Thus, the claim that mindreading and self-consciousness will, according to MRA, emerge simultaneously, appears to depend on the assumption that subjects capable of thinking "She is Ψ" will also be capable of thinking, "I am Ψ". But we have yet to be given any reason to believe that any such subject need be in possession of the first-person concept at all. As such, I take it that the formulation in the main text is to be preferred.

then to a first approximation we should predict that people will also be successful self-attributors" (p.7); *No native control* is the claim that, "to the extent that people are capable of controlling their own mental lives at all, this would be a cobbled together skill that depends on individual and cultural learning" (p.7). I suggest that MRA only predicts *no outstripping*, and that both *equal competence* and *no native control* are only predicted with the help of significant background assumptions. Furthermore, Carruthers et. al. do not offer any direct evidence for *no outstripping*, rather they undermine evidence that might be thought to disconfirm it. As a result, *no outstripping* is not confirmed, it is merely not disconfirmed and, more generally, the contention that, "the predictions made by the third-person-based account of the adaption underlying self-consciousness [MRA] are confirmed" (p.14) remains to be shown.

The mindreading faculty, as Carruthers et. al. understand it, consists in, "a set of representational primitives like THINKS and WANTS, together with some basic inferential principles" (p.2). The idea motivating *no outstripping* is that if self-consciousness is just mindreading that is directed towards oneself, then what one can do for oneself, one will be able to do for others. It seems to me that this is indeed a prediction of MRA. But it is one that is particularly hard to confirm consisting, as it does, in the claim that a certain combination of capacities does not (could not?) exist. Indeed, Carruthers et. al. do not offer any direct evidence that self-conscious abilities cannot outstrip mindreading abilities. Rather, they argue at length (pp.17-25) that certain data from comparative psychology *do not* support, as might be thought, the outstripping of mindreading by self-consciousness. In particular, it does not support the view that some non-human primates have stage-2 self-conscious abilities but only stage-1 third-person mindreading abilities. If this argument is compelling, and minor quibbles aside, I agree that it is, then the *no outstripping* prediction is not disconfirmed by the comparative data in question.

What about *equal competence*? Here I think that it is important to distinguish between abilities, on the one hand, and competence on the other. In the former category we can place the core knowledge that Carruthers et. al. see as at the heart of the mindreading faculty – a set of concepts ('representational primitives') and 'inferential principles' comprising, I take it, a general purpose theory of mind. In the latter, we can place a subject's reliability in correctly attributing psychological states. This latter will, I assume, be sensitive to the sorts of information/evidence available to the mindreading faculty, in a way that conceptual abilities are not. Furthermore, I take it that *equal competence*, framed by Carruthers et. al. in terms of 'success', should be understood as making a claim about competence, not abilities. Now, I take it that the information available for the self-attribution of psychological states differs from that available for their other-attribution. For

example, in the self-attribution of emotion a subject rarely has visual access to their facial expression. Further, something that receives a great deal of attention in (Carruthers 2011), self-attributions of psychological states can rest on an awareness of inner speech, whilst other attributions cannot. Thus, MRA only predicts *equal competence* on the assumption that the information available to the mindreading faculty for other-ascriptions is not (significantly, systematically) inferior to that available for self-ascriptions, nor *vice versa*. This doesn't seem to be a claim that Carruthers et. al. defend. Not here, at least.

Finally, does MRA predict that control of psychological states is not a native competence? I think it is clear that it doesn't. I fail to see how MRA could make any such prediction, since it makes no claims about the nature of metacognitive control. Were there, *pace* the evidence that Carruthers et. al. present, strong reason to accept an adaptation for control, that would in no way disconfirm MRA. It would only do that if the adaptation was for self-consciousness as control regulator. As a result, whilst the fact, if it is one, that there is no adaptation for control of psychological states is evidence against CA, it is not evidence for MRA. Of course, it would be if CA and MRA exhausted the options, but Carruthers et. al. accept that this is not the case (p.2).

The end result is that Carruthers et. al. have identified one prediction made by MRA that, if they are right about uncertainty monitoring and misleading appearances, is not disconfirmed by evidence from comparative psychology. Whilst I haven't here questioned their critical stance towards CA, their positive case made for MRA seems to me to be less compelling than they suggest.

References

Carruthers, P. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford: Oxford University Press.

Goldman, A. 2006. *Simulating Minds: The philosophy, psychology, and neuroscience of mindreading*, New York: Oxford University Press.

Joel Smith

joel.smith@manchester.ac.uk