Making a case for introspection

doi:10.1017/S0140525X0900082X

Alexandra Zinck,[a] Sanne Lodahl,[b] and Chris D. Frith[b,c]

[a]*LWL-Universitätsklinik Bochum der Ruhr-Universität Bochum, Psychiatrie–Psychotherapie–Psychosomatik–Präventivmedizin, Institut für Philosophie, Ruhr-Universität Bochum, 44791 Bochum, Germany;* [b]*Centre of Functionally Integrative Neuroscience (CFIN), Danish National Research Foundation, and Institute of Philosophy and History of Ideas (IFI), and Faculty of Humanities, Aarhus University, Aarhus University Hospital, 8000 Aarhus, Denmark;* [c]*Wellcome Trust Centre for Neuroimaging, University College London, London, WC1N 3BG, United Kingdom.*

alexandra.zinck@rub.de
http://www.ruhr-uni-bochum.de/philosophy/staff/zinck/index.html
sanne@pet.auh.dk www.cfin.au.dk/menu478-en
cfrith@fil.ion.ucl.ac.uk http://www.fil.ion.ucl.ac.uk/Frith/

Abstract: Defending first-person introspective access to own mental states, we argue against Carruthers' claim of mindreading being prior to meta-cognition and for a fundamental difference between how we understand our own and others' mental states. We conclude that a model based on one mechanism but involving two different kinds of access for self and other is sufficient and more consistent with the evidence.

Making a case for introspection. Comparing four different accounts of the relationship between third-person mindreading (meta-representing mental states of others) and first-person metacognition (meta-representing one's own mental states), Carruthers concludes that the capacity to *mindread* is *prior* to metacognition. According to him, basic mindreading is either turned upon others or turned upon ourselves, the latter constituting metacognition. Mindreading is thus the capacity to interpret the other or the self and therefore does not require introspection.

This brings us to the core problem of our critique: Assuming that there is one basic meta-representational mechanism that underlies both understanding the self and other, how can this mechanism be characterized?

In what follows, our analysis hinges on the way Carruthers uses the term *introspection* in relation to basic *mindreading*. Most accounts of mindreading use introspection to describe a special kind of access that we have to ourselves that is not available for third-person mindreading.

Carruthers' account dispenses with this difference of access and the function of introspection. He gives a negative definition of introspection as "any reliable method for forming beliefs about one's own mental states that *is not* self-interpretative and that differs in *kind* from the ways in which we form beliefs about the mental states of other people" (sect. 1.4, para. 3, emphasis in original). Yet, in his architecture of the mind, there is no place for an introspective capacity constituting an immediate and direct inner perception of a belief. This conclusion results from Carruthers' extreme caution about the phenomenology that characterizes introspection and his dismissal of it as misleading.

This thesis of the unreliability of introspection and the necessity to dismiss it as a mode of access to beliefs is supported by data from confabulation and commissurotomy. However, this does not show that one cannot know one's beliefs to be true. We don't necessarily have to concede that beliefs are – or can become – consciously *uninterpreted*; instead we can assume that there are unconscious belief attitudes that can give rise to a conscious event whose content is a belief. In cases of confabulation, this doesn't mean, however, that we are not introspecting this event; it simply means that there is a discrepancy with the underlying belief attitudes.

So, the data suggests that there are mental processes that we are not conscious of. This is not a principled argument against introspective access to our own mental states that is independent of lengthy interpretation. The scope and quality of self-knowledge is limited, whether it is gained by introspection or by self-interpretation. Commissurotomy patients mistake their beliefs for certain actions, yet they do so also under an account of self-interpretation. Self-knowledge and its acquisition by introspection has certainly been overrated in philosophy, but the limitations are equal for self-interpretation.

Aside from this – although he claims that according to his account of mindreading applied to the self, there is "no . . . awareness of one's own propositional attitudes independently of any perceptually accessible cues that provide a basis for self-interpretation" (sect. 1.4, para. 2) – Carruthers does not completely differentiate introspection and self-interpretation according to his mindreading account and does concede there sometimes seem to be introspective qualities during self-interpretation, such as immediacy and effortlessness (sect. 8, para. 3).

So the question still remains of how best to characterize the access we have to ourselves. Contrary to Carruthers, we would like to argue that the immediacy and directness that characterizes

introspection is also present when mindreading others and that this is not a conscious interpretational endeavour. Just as with the perception of the outside world, our brain makes "unconscious inferences" (von Helmholtz 1866) when perceiving ourselves. This is the basis for the experience that introspection is direct and immediate. The same immediacy also occurs when perceiving others (Frith 2007). Nevertheless, there is a difference between the way we meta-represent our own and the mental states of others. When thinking of ourselves, there are more data available, that is, visceral and somaesthetic sensations in addition to a richer knowledge of our own past history. Thus, we are dealing with the same mechanism but with two different modes, one for the self and one for the other. This corresponds to Carruthers' model 2 account.

It accordingly also does not matter whether the mechanism evolved first for understanding others or for understanding oneself. We assume both involve the same underlying mechanism of meta-representation that, endowed with additional sources of information, makes up the different modes of access.

Another point of criticism against a *mindreading is prior* account is that the mechanism of mindreading is third-person directed. Thus, when I direct my mindreading capacity upon myself, I should use a third-person stance. Apart from being an interpretative process, this is also an unnecessarily complex and computationally expensive way of accessing the self. It can be argued that the best explanation is to simply accept the immediate first-person data instead of adopting the complex third-person setup.

A further argument for introspection as a specific mode of access for the self comes from considering why it might be valuable for survival: (1) we can inform others about our reasons for acting in a certain way; (2) we can gain high-level control of our emotion and our behaviour (e.g., Zelazo 2004). Take, for example, a simple learning process. We can learn associations between stimuli even when the stimuli are presented subliminally (i.e., not available to introspection). However, this learning is slow and gradual. If the stimuli are supraliminal, then insightful learning becomes possible through introspection. At some point subjects notice the contingency and will immediately jump to 100% performance (Pessiglione et al., in press).

Contrary to Carruthers, we prefer his model 2 that makes use of one mechanism but involves two different kinds of access: one which is perception-based for interpreting others, and additional introspective access which is available when assessing one's own mental states. Altogether, model 2 is more consistent and parsimonious. It also makes better predictions for pathologies such as autism and schizophrenia in which both kinds of access are impaired.

In sum, this discussion exemplifies that the understanding of how self and other are related is an important topic for research that is generating exciting new empirical and theoretical investigations.

Author's Response