How *Not* to Identify a Research Program Concerning Introspection[*]

Daniel Stoljar, ANU

Abstract: Kammerer and Frankish aim to set out a new research program concerning introspection. I argue they have done no such thing, since the definition they are working with is too general. I further argue that, while it possible to restrict the definition and so formulate a related research program, this will have a different shape to the one they envisage.

Kammerer and Frankish (hereafter 'KF') aim to set out a new research program concerning introspection (Kammerer and Frankish 2023). Introspection, they tell us, may take many different forms. Introspection in artificial agents may be quite different from introspection in non-human animals, and might be different again in humans. They recommend investigating these forms, not simply to guard against over-hasty generalizations from our parochial human case, but to understand that case better.

In some ways I endorse and admire the spirit of KF's paper. I like their open-mindedness and sense of exploration. I do think there are projects in the vicinity of theirs that are interesting and worth pursuing.

Unfortunately, though, I think there is a major problem with their proposal, namely, they fail to properly identify a research program in the first place. The possible forms of introspection, as they define matters, is not a topic you can inquire into. It is too general and too unconstrained. In effect, their project is a 'science of Tuesdays' to borrow the title of one of Jerry Fodor's reviews in the *London Review of Books* (Fodor 2000). You can't have a research program into the possible things that happen on a Tuesday, and the reason isn't that no possible thing happens on a Tuesday. The reason is rather that too many things happen or could happen on a Tuesday, and moreover, these things do not form a natural class. Suppose for example that your hard work pays off and you learn about the ins and outs of 1 or 10 or 100 things that happen on a Tuesday. This would give you no insight into the umpteen other things that happen on a Tuesday, or even the very next thing.

This problem for KF comes very early on in their paper when they provide (p. 3) their provisional definition of what introspection is:

> Introspection is a process by which a cognitive system represents its own current mental states, in a manner that allows the information to be used for online behavioural control.

[*] I'm indebted to a discussion with members of ANU's Philosophy of Mind Work-in-progress group, as well as to Andrew Lee, François Kammerer and Keith Frankish for comments on a previous draft.

In explaining this definition, KF say that 'represents' should be understood 'de re', by which they mean they are interested not simply in the ways in which a cognitive system represents things as mental states but in the ways in which it represents things that are in fact mental states. By 'allows the information to be used for online behavioural control' they mean roughly that the system may at least in principle act on the basis of this information, that it plays or at least could play some specific functional role. So far as I can see, KF don't explicitly say what they mean by 'cognitive system,' but I take them to mean any system or agent that has or could have mental states.

It is certainly a good thing that KF provide a definition of introspection at the beginning of their paper; not everyone does. And by itself it is not a bad thing that the definition, is, as they point out, very liberal. The word 'introspection' can be used in multiple ways; there is no objection to someone defining it in a liberal way if it suits their purpose in doing so. The problem is that the liberality does not suit KF's purpose. You can't operate with this liberal a definition if you aim to identify a research program based on that definition.

To bring this out, consider the following parodies of the notion of introspection as KF understand it:

> Bistrospection is a process by which a cognitive system represents local restaurants, in a manner that allows the information to be used for online behavioural control.
>
> Maestrospection is a process by which a cognitive system represents famous conductors, in a manner that allows the information to be used for online behavioural control.
>
> Quatrospection is a process by which a cognitive system represents (seventies pop icon) Suzy Quatro, in a manner that allows the information to be used for online behavioural control.

Let's concentrate for the moment on the first of these and imagine that two philosophers invite us to join them in a systematic inquiry into the possible forms of bistropsection. "Don't be human-centric", they warn us, "bistrospection may exist in all manner of minds: human and animal, natural and artificial, terrestrial and alien." They go on to add that they don't

mean to restrict themselves to cognitive systems that represent things as restaurants but will also consider cognitive systems that represent things that are in fact restaurants.

We can all agree that something has gone wrong here. It makes no sense to set as your research goal: understanding the possible forms of bistrospection. And the problem again isn't that there are no possible forms of bistropsection. It is rather that there are too many forms, and the forms that there are, or the instances of these forms, don't constitute a natural class. Suppose a group of aliens lands directly in front of *Flavours of India*, and mistakenly represents it as a fuel cache of some rival group. Now contrast that with a different case, for example, when I look up 'restaurants near me' on my phone. Both are cases of bistrospection by the definition. But there is no natural class of cases of bistrospection they fall into. The best you could say is that both fall into the class of representational processes, something which may (*may*) be a natural class. But even this is no help. A research program into bistrospection (or indeed introspection) is not a research program into representational processes as such; it is a research program into a restricted sort of representational process, and the problem is that KF have articulated no restriction that provides a reasonable target of inquiry.

So the definition KF offer of introspection is too liberal, and this means they have not identified a research program as they take themselves to have done. However, as I mentioned, KF themselves note that their definition is liberal. This raises the possibility that what they say in this regard may counter the criticism just set out. Let me therefore turn to the two things they say about this.

The first is that, while the definition is liberal, it doesn't allow in anything at all (pp. 5-6):

> … despite this liberality, not all forms of mental self-representation count as introspective by our definition. If a scientist forms beliefs about their own mental states by applying some scientific theory to themselves on the basis of behavioural data or brain imagery, they are not introspecting. In the current state of technology, such a method would not usually supply information that could be used for online control.

I was baffled by this passage. I certainly agree with the second sentence. It is true that when a scientist forms a belief about their own mental states in the way suggested that wouldn't ordinarily count as introspection. Hence it is understandable that KF intend their definition not to apply to this case.

But the problem is that the definition as stated does apply, or so it seems. Take Jonas Salk, who in developing the vaccine for polio in the 1950s, famously tested it on himself. Salk presumably formed representations on the basis of behavioural evidence of his own physical states, i.e., the ones having to do with polio, and used them for online behavioural control. It is easy to imagine someone like Salk similarly forming representations not of their own physical states but of their own mental states (e.g. unconscious mental states) and using these representations for online behavioural control. There is nothing in KF's definition considered on its own that excludes this as a case of introspection.

In the passage just quoted, it is the last sentence that I think is intended to provide extra material to rule this out. "In the current state of technology," KF say, the scientist in question would not "usually" use the representations for behavioural control. This is where I am baffled. For one thing, what does the current state of technology have to do with it? Surely there are possible minds, including possible human minds, that have no such limitations; and isn't the whole point of KF's paper to urge us to look beyond such limitations? Moreover, and setting aside the current state of technology, KF say only that such representations wouldn't 'usually' be used for online control. What then about unusual cases? The counterpart of Salk's case we imagined might be unusual, but that doesn't stop the definition applying to it, and mistakenly characterizing it as a case of introspection.

There is also a different thing that KF say regarding the liberality of their definition. At several points they consider the possibility of restricting it, and go on to object to this idea on various grounds: in part they think it would overly constrain their project; in part it would run into other problems. To bring out their basic point as I understand it, suppose we restricted the definition by substituting 'phenomenal state' for 'mental state' in the formulation above:

> Introspection is a process by which a cognitive system represents its own current phenomenal states, in a manner that allows the information to be used for online behavioural control.

A problem with this proposal is that, if you are an illusionist, as both Kammerer and Frankish are, there is a way to understand the notion of a phenomenal state according to which no cognitive system is ever in one. Illusionists have problems explaining what this notion of a phenomenal state is, but let us set that aside. It is plain that if by 'phenomenal state' you mean the thing that illusionists think doesn't exist, and if illusionists are right to think this,

there are no processes of introspection on this definition, at any rate no veridical ones. Hence the project of inquiring into possible versions of such processes is either a non-starter or else takes on a disappointingly counterfactual character; neither is good for KF's purposes.

However, while this might be right, it would be a mistake to conclude that it is impossible outright to restrict the definition in a workable way. In fact, there are two reasonable restrictions you could make. The first is suggested by Wikipedia. If you look up 'introspection' on Wikipedia you find the very sensible suggestion that introspection concerns, not mental states as such, but conscious mental states. It is true that in philosophy there are various competing conceptions of a conscious state, but most people including illusionists will allow that there is *an* available conception in which some psychological states are conscious, and some aren't. What this suggests is that we might restrict the definition to conscious states rather than letting it apply to mental states in general.

A second restriction is a standard one in the literature in philosophy on introspection. A major theme in that literature is the contrast between the ways in which we and other cognitive systems represent the conscious states of others and the ways in which we represent our own. When we represent the conscious states of others we tend to use processes that involve ordinary outer perception and certain typical patterns of inference that involve behavioural premises. We can of course represent our own conscious states in that way too, the point rather is that we have available to us in addition a process of representation that is not based on outer perception or inference in this sense. Suppose, in the light of this, we use the word 'distinctive' in this context to mean 'distinct from the kinds of perception or inference we use when representing the conscious states of others;' then we may modify KF's definition a second time so that it is restricted to distinctive processes rather than processes as such.

Putting these two points together, we arrive at this alternative to KF's liberal definition, which nevertheless preserves their vocabulary and structure:

> Introspection is a distinctive process by which a cognitive system represents its own current conscious mental states, in a manner that allows the information to be used for online behavioural control.

There are several noteworthy features of this proposed definition. First, the scientist KF discuss is immediately excluded (as is the counterpart of Salk in our development of their example), and not because of any issue about what usually happens or what the present state

of technology is. It is rather that the scientist doesn't use a distinctive process to form representations about their own conscious states; on the contrary, as KF themselves say, they use behavioural data and brain imagery.

Second, it is a substantive matter whether a cognitive system has a capacity for introspection on this definition. Some philosophers, Gilbert Ryle (1949) for example or perhaps Fred Dretske (2012), deny that humans or perhaps any agent could have such a capacity. On the other hand, that the definition is substantive is no objection. Suppose, to adopt the engineering metaphor that KF employ, we are building an AI system and wonder what it would take to equip the system with a capacity for introspection. This definition provides a clear though challenging answer: we would have to build into it some distinctive process whereby it comes to represent its own conscious mental states.

Finally, while this is a more restricted definition of introspection, it is nevertheless possible to formulate a counterpart of KF's project in terms of it. Different sorts of mind may have different distinctive processes of this kind, and some minds might have more than one. To take one example, consider, as KF do at one point, those philosophers who accept a notion of acquaintance, somewhat in the Russellian sense, and take it to be the basis of a distinctive process of introspection; see, e.g., (Gertler 2011, Gertler 2012). I don't think myself it is plausible that we humans are acquainted with our conscious states in this way (Stoljar 2021). Still, I can imagine a mind that is—a Russellian mind, we might call it. Human minds are not Russellian minds, but it is not impossible for a mind to be Russellian. If so, we have at least two possible forms that introspection can take.

While one can imagine a research project into the possible forms of introspection so defined, however, the shape of this project is very different from the one KF imagine. We may bring this out—and this is the last thing I will do in this commentary—by looking at their remarks about theory of mind, understood as akin to naïve physics or naïve biology, a body of information about how agents in general think, feel and act, that we and other cognitive systems may draw on to interpret others and ourselves. As KF point out, there is an interesting literature in cognitive science, philosophy, and linguistics about the possible forms of theory of mind: what its content and format is, whether it is innate or cultural or some admixture, whether non-human animals have it and if so in what form, etc. They then go on to ask a good question, namely, what the relationship is between studying the possible forms of theory of mind and studying the possible forms of introspection.

Now if you adopt the restricted definition of introspection that we arrived at above, the answer is that these projects are disjoint. In effect the relation between theory of mind and

introspection is on the restricted definition like the relation between theory of mind and perception. Suppose we have some cognitive system that has a capacity for perception, i.e., it has some specific way to represented certain objects and properties in its local environment. The system may also be equipped with a theory of mind. If so, we may ask what kind of perceptual capacity the system has and how this capacity differs from those of other systems. We may also ask what kind of theory of mind it has, and whether this is different from analogous capacities in other creatures. But these questions are separable. Moreover, this remains the case even when we go on to ask, as we should, how the system's theory of mind interacts with its capacity for perception: whether, for example, the theory of mind available to this cognitive system properly characterizes its perceptual capacities or not—both are possibilities after all; or whether the beliefs that the system forms on the basis of perception are influenced by its theory of mind and if so to what extent and in what ways. The same sorts of things apply, on the restricted definition, in the case introspection.

Suppose, however, you adopt the liberal definition of introspection with which KF operate in their paper. Now things look different. Now the system's theory of mind, in the special case in which it is directed at itself, counts as a kind of introspection, even though, curiously, that very same theory of mind when it is directed at someone else, does not. This can make you think that you have discovered a genuinely new form of introspection. You may go on to ask how this new form of compares with other forms. You may even think that the existing literature on the theory of mind can be appealed to as a source of suggestion and example in the course of developing this project.

But the problem here is the one that has been with us all along. The system's theory of mind when it is directed at itself only counts as a form of introspection on the very liberal definition. It is true you could call it 'introspection' if you want to. By the same token, and returning to an example mentioned above, you could call the theory of mind when it is directed at Suzy Quatro, 'quatrospection'. The key point is that in neither case have you identified something that is a proper object of inquiry.

References:

Dretske, F. (2012). Awareness and Authority: Skeptical Doubts about Self-Knowledge. Introspection and Consciousness. D. Smithies and D. Stoljar. New York, Oxford University Press**:** 49-64.

Fodor, J. A. (2000). "A Science of Tuesdays: Review of *The Threefold Cord: Mind, Body and World* by Hilary Putnam." London Review of Books **22**(14).

Gertler, B. (2011). Self-Knowledge. London, Routledge.

Gertler, B. (2012). Renewed Acquaintance. Introspection and Consciousness. D. Smithies and D. Stoljar. New York, Oxford University Press.

Kammerer, F. and K. Frankish (2023). "What forms could introspective systems take? A Research Programme." Journal of Consciousness Studies.

Ryle, G. (1949). The Concept of Mind. Chicago, IL, University of Chicago Press.

Stoljar, D. (2021). "Is There a Persuasive Argument for an Inner Awareness Theory of Consciousness." Erkenntnis.