Peter Carruthers[1] and
Christopher F. Masciari[2]

Subpersonal Introspection

Abstract: *Kammerer and Frankish (this issue) set up a broad tent, intended to encompass all forms of directly-useable self-awareness. But they omit an entire dimension of possibilities by restricting themselves to person-level self-awareness. Their account needs to be enriched to allow at least for model-free meta-representational signals that are not consciously available, but whose appraisal issues in action-tendencies and/or states of person-level emotion.*

Keywords: curiosity; mental effort; meta-representation; model-free; subpersonal.

1. Introduction

Kammerer and Frankish (this issue) — hereafter K&F — offer a broad framework within which a variety of different forms of self-awareness can be organized. They use the term ‘introspection’ to designate the processes that fall within this framework. But they are explicit that they are not attempting to analyse the ordinary meaning of the term. (This is just as well, since many of the processes they describe would almost certainly not qualify as forms of introspection as ordinarily understood.) Instead, they offer a stipulative definition, encompassing any process that produces representations (whether conceptual or non-conceptual) of a cognitive system’s own mental states (whether as such or *de re*), where those representations can

Correspondence:
Email: pcarruth@umd.edu

[1] Department of Philosophy, University of Maryland, USA.

[2] MD Anderson Cancer Center, University of Texas, Houston, USA.

Journal of Consciousness Studies, **30**, No. 9–10, 2023, pp. 75–85
DOI: 10.53765/20512201.30.9.075

influence the online control of behaviour. We are happy to work with this definition in what follows.

Later in their opening section, however, K&F go on to say that they 'assume' that introspection should deliver its outputs at a personal level, in such a way as to be globally accessible for reasoning and decision-making. But it is quite unclear why they make this additional assumption. It is certainly not just an innocent gloss on their initial definition, as we will show. On the contrary, it excludes a range of both actual and possible forms of action guidance via representations of one's own mental states. Perhaps they assume that all actions result from decision-making that involves access-conscious representations. This might be a natural assumption to make from the perspective of common sense. But it is false. In fact, all forms of affective appraisal issue in automatic motor activation, whether expressive (the fear-face, the anger-face, and so on) or instrumental (an inclination to run, in the case of fear; an inclination to attack, in the case of rage). These action-tendencies need to be actively suppressed if they aren't to be carried through to completion. Moreover, there may well be representations of one's own mental states that figure just upstream of the globally accessible states that issue in conscious decisions, too, thereby influencing online action-selection at one remove; indeed, we will see in due course that there are.

K&F's stipulation that the outputs of introspection must be access conscious also seems at odds with their stated motivation to provide a framework for exploring the various *possible* forms that introspection could take, across species and in artificial intelligence. Having given their broad all-encompassing definition, it was surely a mistake for them to then limit the scope of their enquiry by modelling introspection on its most obvious and familiar human varieties, in which people reflectively take decisions on the basis of beliefs about their current mental states.

K&F distinguish three broad dimensions along which kinds of introspection might differ. Introspection can be more or less concept-involving, more or less direct and immediate, and more or less flexible in its internal operations. We suggest that, in restricting introspection to processes whose outputs are globally accessible, K&F have arbitrarily excluded an entire fourth dimension of possibilities. In what follows we will describe two such kinds that are supported by existing evidence, before noting that there might be other forms of subpersonal introspection to be found in other species and that could possibly be built into artificial intelligences.

2. Signals of Ignorance

Carruthers and Williams (2022) attempt to make a plausible case that all forms of investigative behaviour in humans and other animals are directly driven by unconsciously operating signals of ignorance, and by the extent of that ignorance. Forms of behaviour that function to acquire information are extremely widespread across the animal kingdom, from the exploratory flights of bees through to humans asking questions and conducting Google searches. While some kinds of information acquisition might result from an undirected walk through the environment, incidentally scooping up information as one goes, most do not. Most forms of investigative behaviour are targeted and motivated, resulting from failures of recognition or failures of prediction that are appraised as relevant to the animal's goals or needs.

Carruthers and Williams assume that many cognitive processes, and many forms of behaviour, involve competition among noisy neural accumulators (*ibid.*). These race one another to reach a criterion-level set by the context — involving a speed/accuracy trade-off — while inhibiting one another's activity (Usher and McClelland, 2001; Pleskac and Busemeyer, 2010; Forstmann, Ratcliff and Wagenmakers, 2016). When one is required to judge whether the motion in a random-dot stimulus is primarily to the left or the right, for example, activity in motion sensitive regions of the visual cortex builds up over the course of a few hundred milliseconds, with neurons sensitive to rightward motion competing in mutually inhibitory fashion with leftward sensitive ones (subject to random fluctuations that are common to all neural processes). When the activity of either group is high enough, one judges (and responds) accordingly. Many forms of online decision-making have been successfully modelled in these terms (e.g. van den Berg *et al.*, 2016a,b).

Particularly relevant for our purposes is a study by Dufau, Grainger and Ziegler (2012). They successfully used a noisy-competitive-accumulator framework to explain the details of human performance in a word/not-word task. In these experiments, human participants are briefly presented with strings of letters and required to judge whether or not they constitute a real word, responding as fast as they can. Notice that the response 'not a word' can be interpreted as expressing a judgment of 'not known'. For in most iterations of the experiment the strings of letters employed were pseudo-words, and one might think that the only way of judging that they don't constitute a word is through failure to recognize them as such.

In the model tested and confirmed by Dufau and colleagues, activity begins to accumulate in neural populations corresponding to various similar-looking words as each trial begins. Presented with 'housa', for example, accumulators for 'house', 'louse', 'mouse', 'medusa', and others become active. But at the same time, an accumulator that signals 'not a word' also becomes active, competing with and subtracting from the activity of the others. The slower the accumulation of activity in the real-word populations, the faster the accumulation in the 'not a word' population, and vice versa. If the only way of reliably judging that something is not a word is via failures of recognition, then the 'not a word' population will carry the content *not-known* — it is, in effect, a meta-representational signal of ignorance.

Whether or not that is the right way to think about word/not-word tasks, Carruthers and Williams (2022) build on this idea to explain how investigative behaviour is caused. Consider a cat presented with a novel mechanical toy, for example. Neural populations representing various familiar objects will begin accumulating activity, including MOUSE, BALL, STRING, and FOOD BOWL, perhaps. But these compete with a population whose correctness condition is *unknown*. With the latter reaching criterion-level first, innate or previously rewarded investigative behaviour is initiated. The cat approaches the toy, sniffs it, pats it with a paw, and so on. As information about the new entity is gained, reward signals strengthen the likelihood of employing those behaviours in the future, and the strength of the UNKNOWN signal diminishes. Hence the cat gradually loses interest.

Not all failures of recognition or prediction will issue in curiosity, of course, but only when appraised as relevant to stored goals or needs, just as with all other kinds of affective state. In the case of the cat, it may be because mice, in particular, are appraised as relevant, and the mechanical toy is similar enough to being a mouse (evoking enough neural activity) to hold the cat's attention. In other cases, relevance might be provided by a superordinate category such as ANIMATE CREATURE. If one's attention is initially captured by the movement of a novel animal, and one finds learning about animals rewarding, then curiosity will be evoked and sustained.

But why should we think that the neural signal that initiates and sustains investigative behaviour has the (non-conceptual) content *not known*? It is reliably caused in conditions of ignorance, and it is arguably this information carried by the signal, rather than any other, that explains how the downstream use of the signal (causing investigative behaviour when relevant) has become stabilized. Given plausible theories of how content gets fixed, this means that the signal has as its correctness condition that the object or outcome is unknown, or not sufficiently known (Rupert, 2018; Shea, 2018).

Notice that the signals of ignorance postulated by Carruthers and Williams (2022) fit K&F's stipulative definition of introspection. They represent one's own mental state of ignorance (represented *de re* rather than as such), and they initiate and guide investigative behaviour in an online fashion. But they aren't available at the personal level. They aren't themselves accessible in the global workspace. Rather they are received as input and appraised for relevance by the interest/curiosity affective system, with the latter initiating behaviour directly (in the same kind of direct manner as the fear-system generates an urge to run from a fearsome predator). In humans, their effects would manifest in consciousness merely as an urge to look closer, approach, ask a question, or whatever. One is aware of the directly caused motor tendencies whose causes include a signal of ignorance, without being conscious of that signal itself (although when aware of such motor urges, of course, mentalizing humans can reliably report that they are ignorant).

While we currently lack direct evidence of their existence, signals of ignorance are rendered plausible by the likely ubiquity in cognition of processes involving noisy competitive accumulators, and also by the gains in speed and reliability that would attend having a separate accumulator representing *unknown*, rather than allowing investigative behaviour to be caused merely by failures of the object-specific and event-specific accumulators to reach criterion. Indeed, one of the main reasons why mutually inhibitory competitive models have now largely replaced those that postulate an independent race to a decision criterion is that they enable an adaptive trade-off between speed and reliability (Teodorescu and Usher, 2013). Had we adopted K&F's restriction of introspection to person-level representations, we would have missed this important class of forms of self-awareness, or of sensitivity to one's own mental states. But in any case, even if not actual, they surely constitute a *possible* mechanism of introspection according to K&F's broad definition, and are worthy of investigation alongside those that operate at the personal level.

3. Signals of Executive Engagement

Carruthers and Williams (2022) also make a case for another class of unconscious meta-representational signals — in this case, signalling the extent of executive system engagement. Such signals may be somewhat less widespread in the animal kingdom than are signals of ignorance, but are known to exist in rats, at least, and are also likely to exist in some birds. These signals of executive engagement have been investigated and modelled more directly, however. We don't have to rely on general plausibility-arguments as we did in connection with signals of ignorance.

It is well known, of course, that concentrating on a task is generally effortful. Performing a task that requires sustained focused attention is generally aversive (tiring). Attentionally demanding tasks tend to feel like *work*, and people will avoid them by default. However, evaluative learning (specifically, various forms of conditioning) can lead to forms of learned cognitive industriousness, in rats as well as in humans (Eisenburger, 1992; Hosking, Crocker and Winstanley, 2016). Hence people can come to enjoy playing chess or doing crossword puzzles. Indeed, there is now an extensive literature modelling what is called 'the expected value of control', and its role in human and animal decision-making (Kurzban *et al.*, 2013; Shenhav, Cohen and Botvinick; 2016; Shenhav *et al.*, 2017; Winstanley and Floresco, 2016; Inzlicht, Shenhav and Olivola, 2018).

It is well known that choice among actions always involves not just an estimate of the likely value of the outcomes and the likelihoods of achieving them, but also estimates of the physical effort or energetic costs involved (Cisek and Kalaska, 2010; Cos, Duque and Cisek, 2014). But it is now thought that choice among tasks also involves an estimate of the likely *cognitive* effort they require. So one chooses not just on the basis of likelihoods, physical effort, and outcome values, but also taking account of the cognitive effortfulness of the options — how *hard* each task is likely to be, in a cognitive sense. This means that human and (some) animal affective systems must somehow monitor the extent of executive system engagement during tasks, evaluating it negatively by default (prior to evaluative learning), and estimating the costs of such engagement when choosing among future options.

Carruthers and Williams (2022) argue that, for this to be possible, one's affective systems must receive as input a signal or signals that carry information about (and hence represent, given their function) the extent of executive system engagement. The higher the executive demands, the stronger the signals. But like the signals of ignorance discussed previously, signals of cognitive effort are among the inputs to one's affective systems, and hence are not available at a personal level, and don't figure in the contents of the global workspace. What one is aware of is just the output of the appraisal process, which would normally be an aversive feeling attached to the task one is performing or to the task one is considering. Intellectual work generally (but not always) feels bad, just as physical work does. Humans, of course, with their advanced mentalizing abilities, are aware that it is the fact of having to concentrate that makes such tasks feel bad. But this isn't necessary for the system to work as designed, and it seems unlikely that rats have such awareness. Rather, some tasks can just seem bad or unattractive, resulting from an unconscious appraisal of a signal that represents the extent of executive system engagement in those tasks. It is the *tasks* that are represented consciously as aversive, not one's mental effort *per se*.

Signals of attentional engagement, or concentration-to-task, likewise fit K&F's broad stipulative definition of introspection, just as do signals of ignorance. They are non-conceptual/analogue-magnitude representations; they refer (*de re*) to the extent of executive system activity; and they influence online behaviour (albeit not directly, but on the other side of an affective appraisal process). We thus have evidence of a second form of unconscious introspection, or unconscious self-awareness. But it, too, would be excluded from consideration if one were to insist that meta-representational signals that influence online action must also be access conscious to qualify.

4. Other Possibilities

K&F emphasize that one of their goals in offering a broad stipulative definition of introspection is to open up consideration of the *possible* forms that self-awareness might take, both across species and within artificial intelligence. Here, too, it matters that they have closed off a whole dimension of possibilities by insisting that the output of introspective processes must be access conscious, available at the personal level.

To begin this section, we emphasize that information that becomes globally available for use in decision-making and planning is subject to an informational bottleneck (Tombu *et al.*, 2011). Our minds are constantly bombarded with externally caused sensory information as well as information that is endogenously produced (as when one's mind wanders). It is, as a matter of *contingent* fact, impossible for extant organisms to process even a small portion of this information for use in decision-making and planning all at once. To solve the problem, humans and other animals allocate attentional resources to information that is relevant to their current tasks, goals, and values, thus filtering out information that is deemed not relevant. This informational bottleneck is not limited to sensory representations, but also to decision-making itself ('what should I do next?'). There are a limited number of alternatives one can evaluate and select among at any one time. But such bottlenecks are properties of biological minds, resulting, at least in part, from processing that is many orders of magnitude slower than that available to contemporary and future artificial intelligences.

Indeed, even among biological minds there appear to be important differences in cognitive architecture relevant to K&F's assumption of a 'person-level' set of representations. Octopuses, for example, have largely decentralized brains, with separate and sophisticated controllers for each of their arms. Indeed, some suggest that octopuses can make decisions locally rather than relying on feedforward commands initiated by a central processing unit. In a discussion of some findings regarding octopus search behaviour, the authors point out that '[s]earch movements, such as those used by octopuses during exploration and hunting, might require little or no central control and could be performed by local reflexes of the peripheral nervous system (PNS) relying on tactile and chemical information' (Gutnick *et al.*, 2011). Note that if the exploratory behaviour is partly guided by a mental state we might describe as curiosity, and curiosity is meta-representational, then we have an example of a decentralized mind that traffics in non-globally available representations of its own mental states. Put another way, each arm of the octopus might itself be curious. So each would have its own merely locally available representations of its own mental states.

Whether or not octopuses are fully decentralized might be questioned, since presumably the exploratory behaviour of a given limb is restricted — though perhaps only causally and not representationally — by the exploratory behaviour of the others. For example, how one limb explores its environment is constrained by the activities of the others (if one limb explores too far in one direction the others may need to follow). So perhaps some centralized coordination is required. But scepticism about whether or not this is actually what goes on in the case of the octopus is orthogonal to the present discussion. For it is at least possible that coordination is achieved competitively, independently of a central controller. The point is to make vivid a certain possibility that K&F have ignored — that is, the possibility of a highly distributed mind that traffics in merely locally available representations of its own mental states.

How might this work in cases in which coordination among the parts of a system are merely causally coupled? One such process could be something like the winner-take-all model of information selection, of the sort utilized in extant theories of bottom-up attention. Roughly, the winner-take-all model is as follows: different options for information selection are represented locally, and the strongest option wins out and determines the outcome, capturing attention (Koch and Ullman, 1987; Itti, Koch and Niebur, 1998). Crucially, such a process does not require that the various options are represented at the personal level or in some central workspace. Rather, the representations of each possible option — understood in terms of the activity of neural populations — compete with one another, and the population with the highest magnitude (understood in terms of the strength of signal) drives selection. In most creatures, and probably in octopuses too, some centralization is required, but that doesn't strike us as necessary. As Godfrey-Smith (2016) points out:

> To some degree, unity is inevitable in a living agent: an animal is a whole, a physical object keeping itself alive. But in other ways, unity is optional, an achievement, an invention. Bringing experience together — even the deliverances of the two eyes — is something that evolution may or may not do. (p. 87)

Now, if such integration (what is called 'unity' above) is not required by evolution (everything could be based on neural competition, for example), then certainly our design choices need not be so restricted. And since K&F are concerned with *possible* forms of introspection, they need to consider massively decentralized and parallel computational systems.

Here is a toy example: suppose there is some parallel computational system composed of various subsystems and a finite memory cache. Moreover, suppose that the cache is used to store the values of variables utilized by any of the various subsystems when they perform whatever task they were designed for. One possible option would be to have a single representation of the total memory available along with a representation of the memory requirements for all the various subsystems. The latter could then be rank ordered by a central processor in a context-sensitive way — within something like a global workspace — with those higher on the list being allocated more or all of the memory. Once one process is no longer deemed important, the cache would be erased and then allocated to the next process on the list. Another possibility, however, would be to have distinct representations of the memory requirements for tasks of the various subsystems represented *within* each of the subsystems, and have them compete with one another, with the winner being allocated more or all of the memory. One obvious benefit of such a system would be its efficiency and speed since the relative importance of the various tasks wouldn't need to be adjudicated by a central processor (which might be relatively slow and energetically costly). Whether or not this sort of system would *generally* be better or worse than one with a central decision-making workspace is not of concern. We only mention it to show that it is a possible form of self-monitoring that does *not* require centralized (person-level) representations.

5. Conclusion

In this commentary we have argued that K&F have precluded a host of actual and possible forms of introspection by limiting introspection to states that result in person-level representations of one's own mental states. We argued that signals of ignorance and signals of executive engagement are each meta-representational, but neither are available at the personal level. We finally suggested that there are possible forms of introspection that might exist in parallel computational systems but without a central workspace. Since K&F are interested in mapping out the space of *possible* forms of introspection, they ought to drop the requirement that meta-representational states must be cognitively accessible at the personal level to qualify.

References

Carruthers, P. & Williams, D.M. (2022) Model-free metacognition, *Cognition*, **225**, art. 105117.

Cisek, P. & Kalaska, J. (2010) Neural mechanisms for interacting with a world full of action choices, *Annual Review of Neuroscience*, **33**, pp. 269–298.

Cos, I., Duque, J. & Cisek, P. (2014) Rapid prediction of biomechanical costs during action decisions, *Journal of Neurophysiology*, **112**, pp. 1256–1266.

Dufau, S., Grainger, J. & Ziegler, J. (2012) How to say 'no' to a nonword: A leaky competing accumulator model of lexical decision, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**, pp. 1117–1128.

Eisenberger, R. (1992) Learned industriousness, *Psychological Review*, **99**, pp. 248–267.

Forstmann, B., Ratcliff, R. & Wagenmakers, E.-J. (2016) Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions, *Annual Review of Psychology*, **67**, pp. 641–666.

Godfrey-Smith, P. (2016) *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*, New York: Farrar, Straus, and Giroux.

Gutnick, T., Byrne, R., Hochner, B. & Kuba, M. (2011) *Octopus vulgaris* uses visual information to determine the location of its arm, *Current Biology*, **21**, pp. 460–462.

Hosking, J., Crocker, P. & Winstanley, C. (2016) Prefrontal cortical inactivations decrease willingness to expend cognitive effort on a rodent cost/benefit decision-making task, *Cerebral Cortex*, **26**, pp. 1529–1538.

Inzlicht, M., Shenhav, A. & Olivola, C. (2018) The effort paradox: Effort is both costly and valued, *Trends in Cognitive Sciences*, **22**, pp. 337–349.

Itti, L., Koch, C. & Niebur, E. (1998) A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, pp. 1254–1259.

Koch, C. & Ullman, S. (1987) Shifts in selective visual attention: Towards the underlying neural circuitry, in Vaina, L.M. (ed.) *Matters of Intelligence*, Berlin: Springer.

Kurzban, R., Duckworth, A., Kable, J. & Myers, J. (2013) An opportunity cost model of subjective effort and task performance, *Behavioral and Brain Sciences*, **36**, pp. 661–726.

Pleskac, T. & Busemeyer, J. (2010) Two-stage dynamic signal detection: A theory of choice, decision time, and confidence, *Psychological Review*, **117**, pp. 864–901.

Rupert, R. (2018) Representation and mental representation, *Philosophical Explorations*, **21**, pp. 204–225.

Shea, N. (2018) *Representation in Cognitive Science*, Oxford: Oxford University Press.

Shenhav, A., Cohen, J.D. & Botvinick, M. (2016) Dorsal anterior cingulate cortex and the value of control, *Nature Neuroscience*, **19**, pp. 1286–1291.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J.D. & Botvinick, M. (2017) Toward a rational and mechanistic account of mental effort, *Annual Reviews in Neuroscience*, **40**, pp. 99–124.

Teodorescu, A. & Usher, M. (2013) Disentangling decision models: From independence to competition, *Psychological Review*, **120**, pp. 1–38.

Tombu, M., Asplund, C., Dux, P., Godwin, D., Martin, J. & Marois, R. (2011) A unified attentional bottleneck in the human brain, *Proceedings of the National Academy of Sciences*, **108**, pp. 13426–13431.

Usher, M. & McClelland, J. (2001) The time course of perceptual choice: The leaky, competing accumulator model, *Psychological Review*, **108**, pp. 550–592.

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. & Wolpert, D. (2016a) A common mechanism underlies changes of mind about decisions and confidence, *eLife*, **5**, e12192.

van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. & Wolpert, D. (2016b) Confidence is the bridge between multi-stage decisions, *Current Biology*, **26**, pp. 3157–3168.

Winstanley, C. & Floresco, S. (2016) Deciphering decision making: Variation in animal models of effort- and uncertainty-based choice reveals distinct neural circuitries underlying core cognitive processes, *Journal of Neuroscience*, **36**, pp. 12069–12079.