This is a pre-print version. Please refer to the final version, published in the *Journal of Consciousness Studies, 30(9-10), 2023*
https://www.ingentaconnect.com/content/imp/jcs/2023/00000030/f0020009

More possibilities for introspection: Reply to commentators[1]

Authors

François Kammerer (Ruhr-Universität Bochum)

Keith Frankish (University of Sheffield)

Abstract

This paper reflects on and replies to the fifteen contributions (this issue) responding to our target article 'What forms could introspective systems take? A research programme' (also this issue). We focus first on contributions that criticize our research programme, then turn to ones that test our framework against various views and models of human introspection, and finally consider contributions that explore possible variations of introspection in humans, non-human animals, current AI systems, and imaginary minds. We conclude by drawing some lessons for our research programme and making some suggestions for future research on possible forms of introspection.

Introduction

We are grateful to all the contributors for their papers responding to our target article. The article proposed a new research programme focused on possible introspective systems and outlined a conceptual framework for describing such systems, based on the notions of introspective devices and their repertoires. Such proposals invite and require critical scrutiny, both theoretical and practical. Are the proposals coherent and well motivated? Does our

[1] The authors thank the participants in the symposium for their contributions and Graham Horswell at JCS for his patience and his support. They also thank Mike Beran, Alexandre Billon, Heather Browning, Peter Carruthers, Helen De Cruz, Krzysztof Dołęga, Maisy Englund, Philip Goff, Steve Fleming, Robert Long, Pete Mandik, Jennifer Mather, Matthias Michel, Sophie Nelson, Adriana Renero, Eric Schwitzgebel, Daniel Stoljar and Wayne Wu for their comments and advice. Finally, they thank Eileen Frankish for her careful proofreading of the manuscript.

conceptual framework provide a theoretically fruitful way of taxonomizing introspective systems, actual as well as possible? Does the proposed research programme promise to be a fruitful one, which will generate useful insights and hypotheses? These are exactly the questions the fifteen contributions address.

For the purposes of our reply, we shall divide the contributions into three classes. First, we shall consider contributions that question the theoretical basis of our programme, both in general and in specifics. Second, we shall look at contributions that test our conceptual framework by seeing if it offers a useful perspective on existing models of human introspection. Third, we shall look at the numerous contributions that pursue the research agenda we proposed, starting to map the space of possible introspective systems in humans, non-human animals, current AIs, and imaginary minds. Finally, we shall conclude by drawing some lessons from the symposium and anticipating future work in the area.

1. Criticisms of the programme

1.1 General criticisms

Two contributions to this special issue raise general objections to the research programme we propose. We shall start by examining these, before turning to other, more specific objections.

Daniel Stoljar (this volume)[2] raises a radical challenge to our programme. His main concern is that possible introspective systems, in our liberal sense of introspection, are too heterogeneous to form the target of a successful research programme. He compares research on possible forms of introspection to a ‘science of Tuesdays’. Engaging in light parody, he also compares it (among other things) to the study of ‘bistrospection’ — the study of the possible ways in which a cognitive system could represent local restaurants in a manner allowing the information to be used for online behavioural control. Obviously, a science of Tuesdays is hopeless; too many possible things happen on Tuesdays, in ways unrelated to the fact that they happen on a Tuesday, for there to be a possible science of Tuesdays. Similarly, studying bistrospection is unlikely to lead to interesting discoveries, since there are countless ways to represent bistros, and it is doubtful that they have anything interesting in common. If the study of possible forms of introspection is epistemologically akin to a study of bistrospection or a science of Tuesdays, then this is bad news for us.

[2] We shall not repeat this for the other contributions discussed; they all feature in this volume.

Happily, however, there are relevant disanalogies between introspection and, say, bistrospection. Here is the most striking. No one believes that bistros are a robust natural kind, such that representations of them could provide distinctive advantages to a wide range of systems when it comes to behavioural control. On the other hand, many people believe that *mental states* are a natural kind, whose members are unified by (among other things) possession of intentional content and a distinctive role in the guidance of action. And it is prima facie plausible that in acquiring the ability to represent its own mental states, a cognitive system would gain enhanced capacities for self-control. Hence, we should expect many sophisticated cognitive systems to develop introspective capacities, and it should be interesting to ask what forms these systems might take. For similar reasons, we should expect some sophisticated cognitive systems to develop capacities for reasoning about the behaviour of inanimate physical objects, and it might be interesting to explore what forms these capacities could take. We think that introspection in our sense is much closer to reasoning about inanimate physical objects than to Stoljar's bistrospection, and that research on possible introspective systems can be a legitimate part of cognitive science.

Stoljar also insists that our definition of 'introspection' is too liberal to be useful. For instance, we claimed that a scientist who forms beliefs on the basis of brain imagery would not be introspecting because, in the current state of technology, this method would not usually supply information usable for online control. Stoljar is worried, however, that there might be unusual cases where information thus obtained *would be* usable for online behavioural control. And in the future such cases might become common, with artificial implants allowing us to access and use brain imaging data in real time. Stoljar seems to think that it would be preposterous to count cases like these as ones of introspection.

We are happy to bite the bullet Stoljar offers us; in fact, we do not see it as a bullet at all. Representing our own mental states on the basis of brain imagery and using the information for online behavioural control would indeed count as a form of introspection in our book, although a technologically enhanced one. And if the practice were widespread, it would deserve philosophical and scientific attention (the fact that it is not widespread does not make it non-introspective, only less interesting). Of course, one could appeal to semantic considerations to make a distinction between natural forms of introspection and artificially enhanced ones, just as one might distinguish between unaided sensory observation and enhanced forms involving ear trumpets, microscopes, telescopes, night vision glasses, and so on, but we emphasized that we want to avoid such semantic points (see Section 3 of our target paper).

Stoljar also offers a positive proposal, suggesting that a research programme on possible forms of introspection should restrict the concept of introspection to a process that (a) specifically targets conscious mental states (b) is distinct from the processes used to represent the conscious mental states of others. We are not persuaded. Focusing on processes that target only conscious mental states would run counter to our aim of taking a non-anthropocentric view of introspection, and, given the range of conceptions of consciousness available in the literature, would involve a large element of stipulation. And while we did in fact restrict our definition of introspection to processes that generate *personal-level* outputs, even that may be too restrictive, as we discuss in our reply to Carruthers and Masciari below. As for the distinctness restriction, this would not be well suited to exploring non-human mentality. For example, imagine two AI systems, Rob and Bot, who by all available measures are equally good at representing their own mental states. The only difference is that their self-representations employ different processes. Rob represents its own mental states via a sui generis process, while Bot does it via a more general process, which it also uses to represent the mental states of others. Stoljar's requirement entails that Rob introspects, but Bot does not, even though they can know exactly the same things about their own mental lives. Yet there is a highly relevant psychological similarity between Rob and Bot, which Stoljar's definition would erase. Worse: imagine that Rob undergoes an upgrade, which allows it to recruit its existing introspective process to represent mental states of others too, so that the process ceases to be distinctive. Would Rob thus cease to introspect — even though it has merely *gained* a capacity? It is a semantic point of course, but from a non-anthropocentric perspective, it is hard to see what theoretical utility there is in denying that Rob would continue to introspect.

Similarly, suppose that you are studying physical cognition. You might start by defining physical cognition as the representation of features of inanimate objects in a way that enables certain kinds of prediction and explanation. However, in Stoljar's spirit, you might add that physical cognition must rely on some distinctive process. While this would make sense if the goal were to identify something like a physical cognition *module*, for other purposes it might be better to employ a more liberal concept which allows for overlap between the processes involved in representing features of inanimate objects and those involved in representing other features, such as biological or psychological ones. To our mind, the same goes for introspection.

We turn next to **Maja Spener**'s contribution, which also raises a general criticism of our project. Spener focuses on the function of the conceptual framework we propose. She interprets us as aiming to produce a high-level framework for constructing and representing *models* of introspection — a systematic and unified map of the theoretical options available, which allows different models to be compared and contrasted. Spener argues that the construction of such a map is impracticable, if not impossible, since it would have to compare models that are widely heterogenous — targeting different real-world phenomena, abstracting from those phenomena in different ways, employing different methods, and having different theoretical motivations. There is no single theoretical perspective from which a map of such a complex territory can be drawn. So, we cannot deliver what we promised: a way to compare and contrast different possible models of introspection.

Our general response to this is that our aims are more modest than Spener assumes. There are several aspects to this. First, the primary function of our framework is to represent, not possible *models* of introspection, but possible introspective *systems*; it is a framework for introspection itself. And while we hope that the framework will also help theorists compare and contrast different models of introspection, this would be an additional benefit rather than the main aim of the project. Spener may reply that representing phenomena cannot be sharply distinguished from representing models, since phenomena must be represented *via* models. This is true, but there may still be a difference of emphasis between the projects. For example, a phenomena-focused framework wouldn't need to represent different theoretical methods and motivations.

Second, our framework is not intended to encompass all possible theoretical approaches to introspection. Our proposals are addressed primarily to researchers in the cognitive sciences, including empirically minded philosophers, and our framework is designed to represent the *mechanisms* of introspection — the devices involved, their targets, their repertoires, their interconnections, and so on. So, we do not see it as a problem if the framework cannot encompass approaches outside this remit, such as ones that analyse introspection in folk-psychological terms without positing mechanisms. (We mentioned views that treat introspection as a primitive acquaintance relation only as a limit case.)

Third, we did not intend to provide a systematic and unified map of the territory, or even to claim that such a map could eventually be constructed. The map sketched in our paper was a partial one, which will certainly need to be elaborated and might need to be supplemented with other maps drawn to different coordinates (see our talk of a 'first, tentative step'). Thus, while we did aim to produce something more than (in Spener's words) a 'mere list or simple catalogue of theoretical options', we did not aim for a 'systematic representation of the modelling landscape'. What we offered was more like a systematized catalogue, and our sketch map came with an encouragement to produce others.

Finally, the comparison and contrasting of different models of introspection is only one goal of the possible introspections framework, another being an extension of our sense of possibilities — the identification of unexplored regions of theoretical space. This is something we can expect from a systematized catalogue (as opposed to a mere catalogue), and Spener's criticisms do not undermine the heuristic value of our map.

1.2 More specific criticisms

We turn now to two contributions which, while more sympathetic to our overall project, raise objections to specific aspects of it.

Peter Carruthers and Christopher F. Masciari (henceforth C&M) focus on our definition of introspection. We defined introspection as 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. Elaborating on this definition, we added that the introspective information must be *globally accessible* within the cognitive system (at a 'personal' level). That is, introspection must generate metacognitive *beliefs*, or at least metacognitive states that are available for the formation of such beliefs, rather than, say, entirely inaccessible subdoxastic states.

C&M challenge this restriction. They argue that there are possible and actual cases of metacognitive states that are used for online behavioural control without being globally accessible, and that it would be arbitrary to exclude them from a programme like ours, which aims to explore forms of introspection different from the obvious human ones.

C&M describe three types of case. First, they consider subpersonal *signals of ignorance*, which can be modelled using noisy competitive accumulators. These signals do not directly feed into practical reasoning, but they produce feelings of interest or curiosity, which in turn evoke investigative behaviour. Hence, they are representations of current mental states (of ignorance) which are used for online behavioural control, and we should not exclude them from our survey. Indeed, even if such signals do not in fact play this role in humans or other animals, there are certainly *possible* cognitive systems in which they do, and so they fall within our remit.

Second, C&M focus on the better-investigated case of signals of *cognitive engagement*. These are subpersonal representations of the anticipated cognitive effort required by a task. Again, they are not directly accessible at the personal level, but they help shape an overall appraisal of a task (as attractive or aversive, say), which is itself widely accessed, and in this way they play a role in the online control of behaviour.

Third, C&M insist that subpersonal introspection could provide an elegant solution to problems faced by highly decentralized minds, whether natural (such as the minds of octopuses, whose tentacles enjoy a high degree of autonomy) or artificial. They stress that a liberal research programme such as ours, open to non-standard and non-actual forms of introspection, should not exclude such processes from its scope.

We take C&M's challenge seriously. In defence of our original approach, we could note that our working definition is at least partially stipulative. Lines have to be drawn somewhere, and there will always be choices about where to draw them. And while our definition is not heavily constrained by the commonsense conception of introspection, it does retain some elements of it, including the idea that, in introspection, a mind represents *itself*. Subpersonal metacognitive processes, on the other hand, are cases where a *subpart of a mind* represents either some other subpart or the whole mind. So, there is at least a case for restricting the class of introspective processes to personal-level ones.

Ultimately, however, we do not want to rely on this. In contexts like this, pre-existing conceptions and commonsense must give way to theoretical fruitfulness, and we think that C&M make a good case for the view that subpersonal processes could perform what we see as the core function of introspection. Hence, we are inclined to take their point and relax the criterion of personal-level access. Adopting C&M's suggestion, we shall treat degree of accessibility as a new dimension of variation for introspective devices, in addition to the ones we originally proposed — directness, conceptuality, and flexibility. (We discuss this further in Section 4.1 below.)

While we make this concession, we also think that some of C&M's examples deserve more scrutiny. Take the signals produced by a competitive accumulator designed to classify objects of a certain type. C&M take it that a negative classification (the object failing to be classified as a word in the human case, or as a ball in the cat case) can be interpreted as a metacognitive representation of the item as *not known*. This is key to their interpretation of the subpersonal process as introspective. However, the representation could also be interpreted as having a first-order, non-metacognitive content such as *other*. On this view, it would be a sort of hotchpotch-label, characterizing the external object as belonging to some other, unspecified category, or as having an undetermined status, similar to the 'none of the above' response in multiple choice questionnaires. Such an interpretation might better capture the representation's correctness conditions. Take a case where a familiar item is not recognized (say, a cat failing to recognize a ball). Here it is plausible that the competitive accumulator's classification is incorrect (it fails to recognize a familiar object). This in turn implies that the representation's content is better glossed as *other*, rather than *unknown*. Indeed, if its content were *unknown*, then the classifier would never err when it fails to recognize a familiar object, since, in cases of recognition failure, it is *correct* that the nature of the object is currently unknown to the subject, and that it might be learnt with further inquiry. On the other hand, it is not correct that the object is *other* (belonging to another category or having an undetermined status), since it is, in fact, a familiar object. Assuming that recognition failures do indeed involve an incorrect classification, this indicates that the representations in question are best understood as having the content *other*, and therefore as being first-order rather than metacognitive.

However, this worry does not apply to the other cases C&M mention — signals of executive engagement and decentralized minds — and we concede the case for including subpersonal forms of introspection within our research programme.

We turn now to the contribution by **Stephen M. Fleming**. Fleming is supportive of our programme but sees it as most effective when combined with psychophysical methods based on signal detection theory, which have proved so effective in the study of perception. In the perceptual case, subjects try to discriminate some feature of the world, such as the presence or absence of a faint light. Psychophysical analysis is used to control for response bias (the subject's idiosyncratic evidential standard for making a positive response) and so isolate a measure (known as *d'*) of the subject's perceptual sensitivity. In the introspective case, subjects try to discriminate features of their perceptual judgements indicative of how accurate they were. Psychophysical analysis is again used to control for response bias and isolate a measure (*meta-d'*) of their individual introspective sensitivity. Factoring out first-order perceptual sensitivity then yields a measure of introspective efficiency.

This work is still in its infancy, but Fleming argues that it is already shedding light on many of the questions we raise, including the relation between introspection and theory of mind, the number of introspective systems in human and animal minds, the extent of individual and cultural variation in introspection, the effects of meditation, training, and brain damage on introspective capacity, and differences between introspection in biological brains and in AIs. Most importantly, he maintains that an extended programme of psychophysical research on metacognition, starting with the study of simple tasks in laboratory settings, will gradually uncover the computational mechanisms of introspection in humans and other animals, just as similar work on vision has mapped the components of their visual systems.

As we said, Fleming is supportive of our programme, especially our vision of a 'minimal mind', but, as we read him, he is also wary of psychological speculation about introspection that is detached from the precise quantitative foundations provided by psychophysics, and he seems to suggest that the psychophysical approach already provides a unifying approach to the study of introspection, which should be prioritized. (This is why we include him in our 'critics' section.)

In reply, we fully acknowledge the progress of psychophysics and its power to isolate an objective measure of at least some aspects of introspective capacity. It is a hugely valuable tool in the study of metacognition. However, we would also stress the limits of the approach and question its capacity to elucidate all the dimensions of introspection.

First, we worry that the sort of psychophysical approach Fleming outlines will not work for all mental states. We take it that the approach is theoretically attractive precisely because it allows us to measure metacognitive sensitivity (and efficiency) without independent detection of the relevant first-order mental states, thus enabling us to study metacognition without making difficult theoretical decisions regarding, for example, consciousness. But there is a whole range of mental states which do not have a measurable sensitivity, typically because they do not have correctness conditions. This is the case with many conative states, such as desires and intentions. It is unclear how we could determine first-order sensitivity for such states, and, therefore, how we could measure metacognitive efficiency with respect to them, isolating variations in metacognition from variations in the states themselves. Yet conative states are central in our mental lives, and our ability to represent them is central to our introspective capacities. (Some states might be intermediate cases; for example, pain plausibly has a sensory component as well as a conative/affective one.)[3] This looks like a serious limit on Fleming's approach and an obstacle to its serving as a unifying method for the study of introspection.

We are also wary of the idea that the proper way to study introspection is to start with simple metacognitive tasks in artificial settings and gradually move to more complex tasks, following the model set by psychophysical work in vision. While we do not deny the value of such a bottom-up approach, we note that researchers can also study complex introspective processes in naturalistic settings, just as Gestalt psychologists studied the perception of complex features, so that bridges can be built between models of simple and complex introspection. Relatedly, we worry that an exclusive emphasis on quantitative measurement of specific introspective abilities might be limiting, channelling research into the most experimentally accessible paths and concealing options visible only from a wider theoretical perspective. This is one motive for exploring possible forms of introspection; by asking what introspection *could be*, researchers might identify new, experimentally testable models of what it *is*.

It may be that Fleming would not disagree with these cautionary notes, in which case we can all agree that metacognitive psychophysics will play an important role within a pluralistic approach to the study of introspection.

2. Tests of our framework

In this section we consider three contributions which assess our proposals in the light of existing work on the study of human introspection.

Krzysztof Dołęga asks whether our introspective devices framework can be applied to non-philosophical models of human introspection and, if so, whether it offers insights that may drive the development of such models. He focuses on two formal models — Jorge Morales's introspective signal detection theory (iSDT) (Morales, forthcoming) and a metacognitive networks (MN) model developed by Antoine Pasquali and colleagues (Pasquali et al, 2010).

[3] Fleming cites a co-authored article (Beck et al., 2019) which uses his approach to compare metacognition of visual perception, innocuous warm perception, and heat-related pain. However, we suspect that, to make pain amenable to this treatment, he and his co-authors implicitly focus on its sensory dimension — as detection of a certain measurable stimulus (here, noxious heat), which allows the computation of first-order sensitivity, and then metacognitive sensitivity and efficiency. We are doubtful that this captures the conative/affective dimension of pain experience, which seems central to it.

As developed by Morales, iSDT is an account of phenomenal introspection. It assumes that an introspective system monitors experiences in a way analogous to the way perceptual systems monitor the world. Just as an external stimulus produces an internal perceptual response that is the product of both stimulus strength and noise, so an experience produces an internal introspective response (some activation of the introspective system) that is the product of both experience intensity ('mental strength') and noise. And, as with perceptual sensitivity, signal detection theory can be applied to control for the subject's response bias (their individual criterion for making a positive report) and measure their ability to discriminate signal from noise.[4]

The MN model is designed to implement the formation of metacognitive knowledge and its use in post-decision wagering (a betting paradigm used to assess metacognitive awareness in humans). It involves using a second-order neural network to monitor and assess the performance of a first-order network that has been trained to perform a perceptual discrimination task. Crucially, the second-order network does this without access to feedback received by the first-order network, thus enabling it to provide an independent assessment of the first-order network's reliability, which can be used for betting purposes.

Dołęga's aim is not to evaluate these models, but to see whether they can be illuminatingly described within the framework we sketch, which classifies introspective devices along the dimensions of directness, conceptualization, and flexibility. Starting with directness, he notes that the metacognitive networks architecture is somewhat indirect, since the second-order network compares the inputs and outputs of the first-order network and has no direct access to the processing in its hidden layers. Classification of the iSDT model is less clear and depends on whether we focus on the introspective response itself or the introspective judgement about it. Turning to conceptualization, Dołęga notes that both models treat introspection as conceptual, each producing binary outputs that can be interpreted as beliefs. He adds, however, that the iSDT model will require more nuanced description in the light of our distinction between discrimination and characterization. Finally, Dołęga considers flexibility, both as to the target of introspection and the methods employed. He observes that the MN model considered is rigid in both respects (though larger

[4] Morales's use of SDT to *model* pain introspection does not face the objection we raised to the use of SDT to *measure* pain introspection (discussed in our response to Fleming). We just have to suppose that the signal comes from the intensity of the pain experience (whatever that really is) and that pain introspection is sensitive to a mix of signal and noise. The problem arises only when we want to measure pain introspection on the basis of performance on first-order and second-order tasks, without independent access to when pain occurs.

MNs might show more flexibility), but that iSDT is more flexible – for instance, because the placement of the decision criterion seems to be under the subject's control.

Dołęga's assessment is a broadly positive one. He concludes that our framework can be applied to other models and that it can help to drive model development, both by revealing ambiguities in their interpretation (as in the directness of iSDT) and by highlighting new modelling options (for example, regarding the conceptualization of outputs). His main conclusion, though, is that there is a one-many relation between formal models and concrete introspective devices, with models occupying fuzzy regions of the possibility space we defined. He also suggests that our framework might be refined by adding further dimensions, including ones relating to the reliability, precision, and opacity of the introspective process.

We are reassured by Dołęga's results and hope that others will follow his lead in applying our framework and exploring the insights it affords. We also welcome Dołęga's suggestions for expanding our framework, especially as regards reliability and opacity (see Section 4.1 for further discussion of this). Our only substantive comment is that we would have liked to see more discussion of *introspective repertoires*, which provide a further dimension for characterizing models such as MN and iSDT. How does an introspective device group first-order states? How does it characterize them? Dołęga notes this as an avenue for further development, but he does not ask how it applies to the models he considers. At a first pass, the answer seems clearer for the MN model; the second-order network groups first-order states as hits or misses and characterizes them as such. iSDT, on the other hand, seems to be neutral on the matter, though it may not really be, if it is specifically a theory of *phenomenal* introspection. At any rate, this is an issue on which theorists could be more explicit; by leaving repertoire choice implicit or underspecified, they may take some repertoires for granted, inadvertently closing off theoretical options. We hope that the utility of this aspect of our framework will receive further scrutiny.

Adriana Renero's contribution explores the dynamics of introspective episodes. She distinguishes three 'routes' of introspection that cognitive systems can take. Routes are cognitive factors that control the allocation of introspective resources. The *selective route* consists in the top-down selection of specific introspective targets in response to the system's goals, interests, and beliefs. The *cumulative route* involves gathering information about mental states and their sequences, which can be used for purposes of introspective control. Finally, the *predictive route* uses information from past introspective episodes to predict what mental episodes will occur next. Via these routes, which can be refined through practice and instruction, the mind orchestrates its introspective activities using the resources of attention, memory, and inference. Renero supports her analysis by appeal to the reader's own introspective experience, but she suggests that it might be extended by exploring how the routes of introspection vary among humans and how they might be implemented in artificial systems.

We welcome Renero's contribution, which we see as exploring the different ways in which introspective devices can be put to use — and thus as highlighting further dimensions of variation in introspection, reflecting how introspective devices interact with other cognitive resources to construct extended patterns of introspective activity. Any minimally flexible introspective device, coupled with a moderately complex minimal mind (with, say, memory, attention, and some inferential capacities), should be able to adopt basic versions of the three routes Renero outlines. Only a completely inflexible device, whose behaviour was entirely stimulus-driven and whose deployment could not even be influenced by the attentional focus of the wider system, would fail to offer this possibility. From our perspective, what this shows is that the portion of the space of possible introspective devices that corresponds to minimal flexibility is unlikely to be occupied by anything interesting, since — as Renero ably illustrates — it is only through their flexible orchestration that introspective devices can be deployed to serve wider systemic goals.

Wayne Wu explores our framework by using it to situate his taxonomy of (some types of) human introspection. On his view, humans engage in at least three distinct forms of introspection: simple introspection of perceptual experience, introspection of mental action, and complex introspection of phenomenology. Simple introspection redeploys the psychological capacities mobilized in perceptual judgment to introspect perceptual experiences, with merely the addition of 'a concept of experience such as SEEING'. In a somewhat similar manner, introspection of mental action attends to the contents of working memory, giving direct cognitive access to the intentions and plans held there. By contrast, complex introspection of phenomenology is more indirect and less reliable. It may involve simple introspection, but it also employs other processes, including memory, cognitive acts of comparison, and antecedent beliefs, to produce an assessment of how the current experience compares with others.

Drawing on this tripartite analysis, Wu questions our map of possible introspective devices in two ways. First, he notes that the three forms of introspection he analyses are all *agentive*, being mostly under the control of the subject, in contrast to reflex-like forms where aspects of experience force themselves on one's attention. The agentive–reflex dimension of variation corresponds only partially to our flexible–inflexible dimension, suggesting that we need to refine our framework to accommodate it. Second, Wu notes that complex introspection relies *both* on direct informational channels (since it recruits processes involved in simple introspection) and indirect ones (since it appeals to memory of past experiences, antecedent beliefs, and cognitive comparisons). This suggests that our direct–indirect dimension of variation might also need to be modified to capture the number and variety of informational channels mobilized by an introspective device.

We welcome Wu's test of our maps and his suggestions. As we see them, the dimensions of variation we proposed (direct–indirect, flexible–inflexible, conceptual–non-conceptual) can all be broken down into multiple subdimensions, and the choice of which dimensions to centre in any given case may depend on theoretical context (partially addressing Spener's concerns discussed above). And, in many contexts, Wu's suggestions may be appropriate. In line with his first suggestion, the flexible–inflexible dimension might be broken down into two subdimensions: (a) agentive–reflex, reflecting how far the activation and targeting of the device is under agential control, and (b) modifiable–non-modifiable, reflecting how far the internal functioning of the device can be modified in response to other cognitive factors.

As for what Wu calls 'complex introspection', which uses both direct and indirect informational channels, we think this might be best characterized in our framework as recruiting various introspective devices, some direct and some indirect. Still, it hints at the possibility that a single device might mobilize various informational channels with different degrees of causal directness — a possibility that indicates the need to refine our framework. We favour an option on which the directness of a given device can be specified, first by counting the distinct informational channels involved in the production of the device's outputs, and then by providing, for each of these channels, a characterization of its causal directness. This is just a first pass, however, and the framework might need to incorporate further factors, including a measure of how information from multiple channels is integrated and articulated. Note also that this requires a distinction between the number of *channels* mobilized by a given introspective *device*, and the number of *devices* employed by a given introspective *system* (the latter already encompassed in our original dimension of the *unity* of introspective systems). This distinction might not always be easy to draw.

Though we embrace these two suggestions, we want to scrutinize some of Wu's other claims. We shall make two points. First, Wu claims that simple introspection is a mere redeployment of perceptual capacities with the addition of a concept of experience, such as SEEING. As Wu puts it, simple introspection is perception with a ‘small wrinkle’. However, this ‘small wrinkle’ might be more complex than it seems. First, it presupposes that the introspecting subject possesses concepts of experience, such as SEEING, and we might wonder about the nature and provenance of the introspective repertoire to which these concepts belong. More generally, we feel that Wu’s contribution could be interestingly extended through an investigation of the repertoires of introspection, which are left largely uninterrogated.

Second, the ability to redeploy perceptual capacities and the possession of concepts of experience are not jointly sufficient to achieve simple introspection. Also needed is a capacity to apply the *right* concepts (EXPERIENCE, SEEING, HEARING, etc.) to the *right* informational contents. To make this clear, imagine a case where you see a box which looks red but which you know to be in fact white (suppose you know you are wearing red-coloured lenses). Arguably, the contents *there is a red box* and *there is a white box* are both available in your cognitive system. The challenge then is to correctly apply the right concept in the right way, so as to output that you SEE (have a visual experience as of)[5] a red box but not a white box, since you believe, but do not visually experience, that there is a white box. Or, alternatively, imagine a case where you visually experience the presence of an object on your right (without identifying it; imagine the object is at the edge of your visual field and seen as a mere *something*) while auditorily experiencing the presence of an object above you (again, without identifying it as more than a *something*). The contents *there is something on my right* and *there is something above me* are both available in your cognitive system, and the challenge for simple introspection is to correctly output that you SEE an object on your right and HEAR an object above you, rather than the other way around.

Depending on the nature of your introspective devices, successfully introspecting in these cases may take a lot more than a small wrinkle. On the one hand, you might have a range of specialized devices which take up the outputs of each of your perceptual subsystems and add the right concepts to them (SEEING, HEARING, etc.). In such a case, correct applications of the concepts would be trivial; there is a device that is hardwired to apply SEEING to the outputs of your visual subsystem and to no others. On the other hand, you might have a general device that takes input from all active informational contents and uses various clues

[5] We treat the concept SEE, mentioned by Wu, as a concept of visual experience, not of factive seeing, since the subject would arguably admit they were factively seeing a white box.

to determine what is visually experienced, what is auditorily experienced, what is believed, and so on. Such a device would probably need to use a diversity of informational channels, including ones providing information about the environment, and would thus be more indirect than the specialist devices.

We do not take a stance here on which sort of devices human 'simple' introspection actually uses (though the fact that human subjects sometimes commit what early-twentieth-century introspectionist psychologists called the 'stimulus error', confusing properties of the stimulus with properties of perceptual experience, suggests that we do not have specialist devices[6]). Moreover, the architectures described do not exhaust the possibilities. (Another option would be for the outputs of various subsystems to be tagged with identifying markers, which are then read by a general introspective device.) Our point is that even an apparently simple form of introspection can be performed by devices of different kinds (more or less specialized, more or less direct, etc.) with different advantages and disadvantages (specialized devices might be more reliable, but a suite of such devices would be more cumbersome and costly than a single general one).

3. Reflections on cases

In this section we shall consider the contributions that take up our challenge by exploring varieties of introspection that might be found in humans, non-human animals, current AIs, and imaginary minds.

3.1 Humans

Alexandre Billon dedicates his piece to variations in introspection among subjects with certain mental disorders, specifically schizophrenia and depersonalization disorder. These subjects tend to report that their experiences lack certain theoretically contested features, such as 'mineness', feeling of reality, and phenomenality. It could be that their experiences really are impoverished in this way relative to those of neurotypicals, but another option is that these subjects actually have superior introspective abilities, which reveal the absence of

[6] The thought is that stimulus errors are easier to explain if there is one general introspective device which draws on various information sources to decide what is merely experienced as opposed to what really exists, than if we have specialized introspective devices that directly apply experiential concepts to the outputs of sensory subsystems.

features that neurotypicals mistakenly assume their experiences to possess. That is, they might suffer from what Billon calls *superintrospectionitis*.

Billon doubts the superintrospectionitis thesis and offers a ‘fine-tuning’ argument against it. Our minds (including our introspective capacities) are fine-tuned, fragile systems, and we should not expect disorders of such systems to enhance their performance. Billon concedes that this is not absolutely conclusive. For example, introspection might be *plurimodal*, consisting of distinct, parallel components performing distinct functions, and if it is, then a breakdown of one of the components might evoke compensating overperformance from the others. However, Billon thinks that his argument puts the burden of proof on proponents of the superintrospectionitis thesis. He also considers the suggestion that meditation and the phenomenological reduction confirm that the introspective reports of patients with schizophrenia and depersonalization disorder are more accurate than those of neurotypicals, but argues that the evidence is inconclusive. He concludes by rejecting the superintrospectionitis thesis.

In the course of his article, Billon highlights what he takes to be gaps in our target article. For instance, when considering the hypothesis that introspection is plurimodal and involves ‘various independent, parallel processes’, he notes that this ‘important aspect of the architecture of introspection’ is ‘not mentioned in [our] framework’. However, this is not entirely correct. It is true that we did not use the term ‘plurimodal’, but we did mention *unity* as a dimension of variation in introspection. This dimension, which integrates various factors including the diversity of introspective devices and their degree of coordination, allows the possibility that introspection might be realized by (as we put it) ‘many uncoordinated devices with diverse repertoires’ — something close to Billon’s plurimodal form of introspection.

Elsewhere in his article, Billon criticizes the hypothesis (which we examined but did not endorse) that meditation could afford a highly flexible sort of introspection. In response, he distinguishes two sorts of flexibility: use-flexibility and tool-flexibility, the latter deeper than the former. A system is use-flexible if it will produce abnormal outputs when fed abnormal inputs; a system is tool-flexible if can be modified to produce abnormal outputs even when fed normal inputs. Billon stresses that use-flexibility is widespread; even our visual processes, which we took as examples of highly inflexible processes, are use-flexible, since they will produce abnormal outputs when fed abnormal inputs (for example, contraction of the ciliary muscle causes temporarily short-sightedness). And Billon insists that there is no reason to believe that meditation reveals anything deeper than use-flexibility in introspection.

We grant that the fact that meditating subjects introspect differently *while meditating* does not in itself establish that introspection is tool-flexible, since the subjects are, arguably, feeding their introspective systems with abnormal inputs. But that doesn't settle the question of tool-flexibility. The issue is whether meditative *training* has effects on introspection, both that performed during meditation and that performed outside it. And there are prima facie reasons to think it does. In many contemplative traditions, expert meditators evoke 'stages' of the meditative processes that can only be attained through long practice of meditation. One well-known example lies in the numerous 'progressively deeper stages of meditative concentration (*jhana*)' described in the Pali Canon of the Theravadin Buddhists (Katz, 2016). This suggests that frequent meditation influences introspection in meditation. Moreover, Billon himself mentions evidence that meditative training influences introspection outside meditation (Baird et al., 2014), as trained meditators perform better than non-meditators on memory-related metacognitive tasks. We do not take a stance on this issue, but we believe it is an open possibility that introspection possesses a tool-flexibility which meditation can exploit and manifest.

As regards Billon's general argument, we are unsure how much the fine-tuning argument shows. Billon seems to ignore potentially relevant evidence here — for instance, the well-documented fact that some other serious mental disorders can be accompanied with 'islands of genius', as in the case of savant syndrome. Savant syndrome is uncommon but not exceptional: it is claimed that as many as one in ten persons with autistic disorder (about 50% of savants are autistic subjects) have 'such remarkable abilities in varying degrees' (Treffert, 2009, p. 1351).[7] But if the fine-tuning argument were sound, savant syndrome should be extraordinarily rare. The relatively widespread nature of savant syndrome certainly does not show that that the superintrospectionitis thesis is true, but it casts doubt on the dialectical power of the fine-tuning argument.

We ourselves do not find the superintrospectionitis thesis especially attractive, but we believe it is important to investigate potential differences in introspection in various forms of mental divergence and disorder. Moreover, we believe that this investigation could benefit from a suspension of questions about the *value* of any differences identified. Instead of asking whether disordered or neurodivergent subjects are *better* or *worse* introspectors, it

[7] The 10% estimate is the 'generally accepted figure in autistic disorder', but some estimates put the proportion of savants among autistic subjects much lower — as low as 'one or two in 200' (Treffert, 2009, p. 1352).

might be more fruitful to focus on identifying exactly *how* they introspect. How do their introspective reports differ from those of neurotypical subjects and what hypotheses regarding introspective devices, general-reasoning capacities, and so on might explain this variation?

Bryce Huebner and Sonam Kachru (henceforth H&K) explore the ways introspection — notably as performed during meditation — has been conceived within the Buddhist tradition. They explore three conceptions of introspecting and meditating minds. First, there is the *Entirely Mindful Observer* (EMO), who at each moment directly represents their current mental state and nothing else. An EMO registers only momentary mental events and has no way of tracking patterns in experience or of inferring the theoretical and practical significance of mental episodes. Second, there is the *thin mind*, which centres the view developed by Digṅāga, on which experience involves a pre-conceptual form of reflexive self-awareness, unaccompanied by an observing self.

While H&K acknowledge that the experiences of some meditators seem to conform to one of these two conceptions, they argue that meditation in fact requires a much more active involvement and shaping of experiences by the meditating subject. They favour a *thicker mind* conception inspired by early Yogācāra thinkers, which does not identify mentality with self-awareness and phenomenality and sees minds as collections of looping impersonal cognitive and evaluative processes interacting in complex ways with the environment and continually shaped by an array of metacognitive processes. On this view, the meditator cultivates new levels of metacognitive control, selectively attending to aspects of these looping processes and actively shaping how they unfold in normative ways, thus sculpting experience itself.

If our minds do not ordinarily seem this way, this is due to our undergoing a distorting process of ‘afflicted mentation’, which creates the appearance of a unified, well-ordered internal experiential domain belonging to a ‘subject’ and independent of impersonal interactive processes. Meditation can be seen as an attempt to disturb afflicted mentation and achieve new perspectives on our own minds. Through the use of metaphors, such as seeing our experiences as a magician’s illusion or thinking of ourselves as automata, we can try to ‘reconfigure our introspective repertoires’. But the goal of these practices is not to ‘make access to mental states more certain’, but rather to cultivate ‘skills for navigating the complexities of a mind in motion, in ways that are not unlike learning your way around a city’.

H&K’s contribution is highly useful. To begin with, it is a corrective to simplistic conceptions of Buddhist thinking about minds and meditation, emphasizing the rich diversity of approaches available within the tradition. Moreover, the thicker mind conception H&K defend is fertile ground for exploration of variations in human introspection.

First, it gives a key role to afflicted mentation, a process which introspectively foregrounds a narrow range of interacting impersonal processes and ignores a mass of others, thus producing errors with both an epistemic and an ethical dimension. This raises the possibility that typical introspection produces partially arbitrary — and modifiable — groupings, as well as (possibly adaptive) forms of mischaracterization. We think this idea should be key to research on possible forms of introspection, particularly explorations of introspective repertoires. When understanding how introspection groups and characterizes mental states, we must keep in mind that its groupings can be more or less natural and its characterizations more or less accurate. We both have defended illusionism about phenomenal consciousness (Frankish, 2016; Kammerer, 2021), and we believe that introspection mischaracterizes some mental states as phenomenal. Neither illusionism nor even belief in introspective mischaracterization is required to motivate an exploration of possible forms of introspection, but those views can certainly contribute to the motivation.

Second, one particularly interesting idea we take from H&K’s contribution is that different introspective repertoires may serve different *roles*. For example, ordinary afflicted mentation creates and preserves an inflated sense of self, which might be adaptive. Reconfiguring our repertoires to disturb afflicted mentation, on the other hand, might serve the therapeutic aim of diminishing discomfort and disquiet. This also highlights the diverse ways in which introspective capacities can be linked to behavioural modulation. In particular, it reminds us that introspection does not merely help us predict our own behaviour, but also helps us form a conception of ourselves — of our nature, goals, beliefs, and values — which in turn shapes who we are and how we interact with our environment. Research on possible forms of introspection should recognize the variety of functions introspection can serve and the potential tensions between them. These functions include adjusting behaviour to our goals via accurate self-prediction, maintaining a robust sense of self through change, gaining knowledge of one’s psychological dynamics, shaping those dynamics through attentive metacognitive activity, reaching equanimity and calm, and avoiding privileging oneself in ethically dubious ways.

3.2 Animals

Heather Browning and Walter Veit (henceforth B&V) endorse our research programme and take up our challenge to explore the diversity of introspection in non-human animals ('animals' for short) and AIs. They argue that, despite the theoretical challenges, research into animal introspection is feasible, and they suggest that it can be studied through direct self-report methods, teaching animals to use symbols or produce other behaviours to indicate their mental state. They also discuss the use of indirect self-report methods, which involve inferring introspective processes from an animal's behaviour, such as choices that reflect its confidence in its own judgements. Such indirect methods are pervasive in the field of animal metacognition (see also the contribution by Englund and Beran discussed next).

Although B&V acknowledge the difficulties inherent in applying these approaches to animals, they argue that the results can be made more robust through multiple tests across different modalities. They also argue that the study of animal introspection can be enriched by adopting an evolutionary perspective and thinking about how animals need to respond to various degrees of *pathological complexity* (or life-history complexity) — the challenge of finding the right strategies to maximize their fitness, given their ecological niche and stage in their life cycle. They suggest, for example, that animals with complex social lives, such as corvids and apes, are likely to have evolved introspective capabilities, since these would enable them to use mentalistic self-modelling to predict and manipulate each other's behaviour. Finally, B&V extend their insights to the study of introspection in AIs, arguing that considerations of life-history complexity can help us understand the kinds of introspection that would be useful for artificial creatures, such as social robots.

We appreciate B&V's endorsement of our approach and agree with many of the points they make, especially the need to consider life-history complexity when assessing the introspective abilities of animals. In response, we wish to make two points, one relating to self-reports in animals and the other to the application of life-history complexity considerations to AIs. We shall make the first point here, saving the second for the next section.

It is widely assumed that animals lack the ability to make verbal reports and that this presents a challenge to the study of animal minds — and thus of animal introspection. We cannot get animals to tell us about their knowledge, desires, feelings, and so on. However, B&V propose that direct report methods may be suitable for animals and could be used in the study of introspection. They note that some animals, including apes, parrots, and dogs, have been trained to use symbols, and they argue that we should therefore be able to teach them to report their own internal states. Indeed, they note that this may have already been accomplished with dogs, citing Franck Péron (Péron, 2012), who himself refers to a study by Alexandre Rossi and César Ades (Rossi & Ades, 2008). (B&V go on to note that non-symbolic direct self-reports might also be possible, but we can set that aside here.)

This is fascinating, but we remain skeptical at this stage. As B&V seem to concede, an animal report could be interpreted as expressing either a first-order mental state or a second-order (introspective) state representing a first-order one (they say that an animal's report might concern 'first-order experiences or metarepresentations'). In the dog study they cite, the dog's use of a certain symbol could be interpreted as expressing either a desire for food ('Give me food now') or an awareness of a desire for food ('I currently desire food'). The report is introspective only on the second interpretation. Disentangling these two interpretations can be challenging with human reports, and the difficulty is much greater with animal ones, where verbal clarification is unavailable. Moreover, the first, non-introspective interpretation, is the simplest and should likely be the default one; in the absence of compelling evidence, we should assume that the dog is using the symbol to express, rather than describe, its desire. Having said this, we do not rule out the possibility that dogs and other animals can use symbols to make genuine introspective reports, but rigorous further work will be needed to confirm it.

Maisy D. Englund and Michael J. Beran (henceforth E&B) also support our programme, and they offer a perspective from comparative cognitive science, focusing on studies of metacognition in non-human primates. Reviewing recent research, including data from uncertainty monitoring studies, they conclude that there is compelling evidence that non-human primates have some awareness of their own mental states and can use this awareness to modulate their behaviour. They argue that the evidence from these studies is not explainable by associative learning alone, and they cite findings from self-control studies with chimpanzees, which they believe provides some of the strongest evidence for mental self-monitoring. E&B conclude that some non-human primates have metacognitive processes that are *introspection-like* — that target their own mental states and feed into online behavioural control. However, they hesitate to label these processes *introspective* in our sense, since they are uncertain that they are genuinely representational. E&B conclude their paper with some recommendations for future research on animal introspection from a comparative perspective, proposing that researchers operate with the default assumption that animal introspective devices are (in our terms) both *direct* and *non-conceptual*. They argue that this is the most parsimonious assumption, and therefore the most plausible.

We concur with much of what E&B say, and we are happy to see that comparative psychologists take the possibility of animal introspection seriously and endorse its systematic investigation. We have two points to make in response. The first concerns E&B's hesitance to employ our notion of introspection; the second concerns their recommendations for future study of animal introspection.

First, E&B hesitate to label the metacognitive processes they discuss *introspective*, since they are uncertain that they are genuinely representational. This uncertainty appears to hinge on an implicit adoption of a heavy-duty concept of representation, which takes representation to require explicit conceptual characterization and possibly a capacity for linguistic articulation. However, when we defined introspection as representational, we used an explicitly *inclusive* sense of representation (see Section 1.1 of our paper). Given this, we feel there is no substantive disagreement here, and that the processes E&B discuss are ones we would accept as introspective. Indeed, this is the conclusion E&B end up drawing, noting that, given our definition, the ‘introspection-like’ compromise may not be necessary and that the processes in question ‘can be called introspective even though they may be distinct from the processes of human introspection’.

Second, we wonder if the assumption of the directness of animal introspection is as well justified as E&B claim. One reason to qualify it is that some non-human animals plausibly possess a rudimentary ‘Theory of Mind’ (see our target paper, Section 4.2.2). It is not implausible that chimpanzees, for example, use mentalizing capacities to predict the behaviour of others, modifying their own behaviour in the light of their predictions. But if they can do this, then perhaps they can apply their mentalizing skills to themselves, forming representations of their own mental states that can be used for online behavioural control, and so achieving introspection. This is not an outlandish hypothesis. Some theorists hold that human introspection depends heavily (if not exclusively) on a self-applied theory of mind (see the map of possible introspective devices in our target paper). (This need not involve explicit representation of folk-psychological principles; it might simply involve the self-application of implicit mentalizing capacities.) We do not feel tied to this particular theory of human introspection, but it is an interesting possibility, and we fail to see why the same possibility should be dismissed at the outset when it comes to non-human animals such as chimpanzees, who are plausibly capable of basic mindreading and whose representational repertoire might therefore already includes representations of mental states. But self-applied theory of mind would count as an *indirect* introspective device, since its main inputs would be information about the subject's behaviour or environment.

That some non-human animals may possess indirect introspective devices does not, of course, imply that they lack direct ones. Indeed, direct introspective devices may be far more common in the animal world. Moreover, it may be that the most researched capacities in the field of animal metacognition, such as uncertainty monitoring, depend on direct introspective devices. Still, the assumption that animal introspection is direct should be approached with caution and probably rejected if understood in an exclusive sense.

Jennifer Mather and Michaella Andrade (henceforth M&A) consider the possibility of studying introspection in cephalopods. They are skeptical, arguing that we must first ascertain whether cephalopods have mental states at all — a matter which, they claim, is not yet settled. They collate behavioural evidence suggesting that species of cuttlefish, octopuses, and squid manifest mental states, including perceptions, beliefs, and intentions. Focusing on the complex tactics of sexual display in cuttlefish and squid, which they interpret as indicative of intentions, M&A speculate that this behaviour might even require introspective processes.

We feel there is a tension in M&A's perspective. On the one hand, M&A are skeptical about research on cephalopod introspection on the grounds that we cannot assume that cephalopods have first-order mental states to introspect. Yet, simultaneously, they present copious evidence that cephalopods *do* possess such states, including sophisticated beliefs and intentions. In fact, M&A appear inclined to endorse a robustly mentalistic interpretation of the behavioural evidence, in contrast to deflationary interpretations in terms of merely behavioural regularities. From their own standpoint, this would position cephalopods as promising candidates for possession of introspective capacities. Thus, we fail to see why cephalopods should be excluded from research into animal introspection, of the sort advocated by Browning and Veit in their contribution and by Englund and Beran in theirs.

A similar tension is observable in the detail of M&A's argument. For example, when discussing certain octopus behaviours that justify attributions of first-order mental states, they hypothesize that the animal's actions are motivated by an awareness of an informational deficit ('clearly the octopus had decided that it needs more information, and investigates to acquire it'). This naturally invites an introspective interpretation; the animal is sensitive to its own mental states (its lack of certain information) in a manner that allows for online behavioural control. And, while it is possible to interpret curiosity behaviour as motivated by first-order states only, a metacognitive/introspective hypothesis is at least worth taking seriously. As we have seen, Carruthers and Masciari consider the hypothesis — with respect to the curiosity of octopuses' *arms* — in the course of arguing for the possibility of decentralized, subpersonal introspection. Setting aside the question of subpersonal introspection, we think that the curiosity displayed by octopuses could plausibly be seen as an indication of cephalopod introspection — a possibility M&A may have overlooked.

3.3 Artificial intelligences

Artificial intelligence is an important arena for research on possible forms of introspection. It is here that we may expect to find forms of introspection that diverge most radically from human ones, and in our target article we identified AI as a primary field for case-driven exploration of possible introspective systems. Numerous articles in this symposium engage in some way with the introspection in AIs, but the contribution that devotes most attention to the topic is the one by **Robert Long**. (Here we focus on discussions of actual AIs; we shall discuss contributions that explore fictional artificial minds in the next section.)

Long's contribution bears on the introspective capabilities of large language models (LLMs), such as the now ubiquitous ChatGPT, which has propelled LLMs into the everyday experience of millions of people. While LLMs demonstrate remarkable linguistic abilities, they appear ill-equipped for introspection, and some theorists doubt they even possess representations in any interesting sense (e.g., Bender et al., 2021; Marcus & Davis, 2020). Nonetheless, Long contends that LLMs could, in principle, develop a form of introspection. He argues that they may already possess genuine representations of mental states, enabling them to model and 'understand' these states in a non-trivial sense. And he suggests that some may even exhibit rudimentary ('proto') introspective abilities, such as the capacity to assess their own knowledge or certainty about things. Long also proposes that we might cultivate more robust introspective capacities in LLMs by training them to answer questions about themselves while 'cleaning' their training data of material that would enable them to respond in a parrot-like way. Finally, he notes the potential moral importance of LLM introspection, suggesting that it provides independent information about their possession of morally significant mental features.

There is much to discuss in Long's contribution. One point concerns his recommendation that we clean the data used to train LLMs to introspect. The rationale is to ensure that any apparent introspective reports the LLMs make are the result of genuine introspection rather than mere recombinations ('back-parroting') of introspective statements found in the training data. However (as Long notes), if we were to eliminate all mental terms from the training data, LLMs would fail to generate any mentalistic reports at all. Long hopes for a sweet spot between these two extremes, while granting that it may not exist.

In response, we note that, although humans now learn mentalistic terms through linguistic immersion, there must have been a point in history when our ancestors first originated these terms. We can imagine that an LLM, when exposed to extensive datasets composed of purely behaviouristic statements, might originate representations, not only of non-mental entities, but also of mental ones, precisely as effective strategies for predicting diverse behaviouristic datapoints (somewhat like Sellars' imaginary character Jones, who devises a theory of mental states to explain human behaviour (Sellars, 1956)). If this were to happen, it would certainly dispel any doubts about data contamination and back-parroting of introspective statements.

There could also be less drastic solutions to the problem of data contamination. For instance, we could remove all *introspective* uses of mental-state terms from the training data but retain non-introspective ones, which reflect the explanatory and predictive role of mental states, as in the Lewisian list of platitudes supposedly constitutive of our own theory of mind. Or perhaps, instead of *removing* introspective statements from the training data, we could *add in* a proportion of pseudo-introspective statements employing fake mental concepts with fanciful functional roles, and see if the LLM learnt to employ the genuine mental concepts rather than the fake ones when asked about its own internal states. (Here an introspective capability would be revealed not only by the production of certain statements, but also by the *non-production* of others.)

Another point concerns Long's claim that AI introspection may have moral significance as well philosophical and scientific interest, since it gives us reason to believe that AIs possess morally significant mental states, such as desires or phenomenally conscious states. The idea is attractive, suggesting a way of bypassing contentious theoretical debates about the nature of these states in humans and the difficulties inherent in extrapolating from theories of human mentality to claims about AI minds. However, even assuming the possibility of reliable AI introspection, this approach might not avoid contentious theorizing. There is no clear consensus on what exactly makes mental features such as consciousness or desire morally significant. Even if an AI’s introspective reports could convince us that it exhibits some versions of these states, we might still question whether it has *morally significant* versions, rather than impoverished, morally neutral ones. A theoretical decision regarding what makes these features morally relevant in the human case would be still required. But then the ethical significance of direct introspective access to AIs' mental states is limited.

One reason why the possibility of LLM introspection is particularly fascinating is that the task of LLMs — roughly, next-word prediction — does not seem to require introspection. Of course, the same could be said about any representation (or ‘world-model’) of non-linguistic reality. Prima facie, no representation of the world beyond language is required for next-word prediction. However, given that LLMs do appear to possess world-models, we can conjecture that modelling the world is simply an efficient way to predict the next words in a massive corpus, since the world was causally involved in creating the corpus. Still, this hypothesis does not straightforwardly extend to mentalistic self-modelling by LLMs (as opposed, perhaps, to mentalistic modelling of humans by LLMs). Why would an ability to represent *one’s own current internal states* be an efficient way of predicting the next word in a massive corpus whose production was not causally influenced by those states? This might change somewhat in the future, if current LLMs significantly shape the expanded corpus used to train their descendants, either by directly contributing to it or by shaping the world described in it. Even then, though, it would not be the current internal states of the introspecting LLM that had produced the data to be predicted, and the utility of an LLM extending its world-modelling to introspection would depend on how similar its own internal dynamics were to those of the past LLMs that had influenced the data.

There may be a lesson here for the application of ‘life-history complexity’ considerations to AIs, as recommended by **Browning and Veit** (B&V) (see above). B&V suggest that focusing on the tasks and needs of AIs — and the aims of their designers — should shed light on the forms of introspection they are likely to possess. From such a perspective, it looks unlikely that LLMs will develop introspective capacities, given their impoverished function and life history and the causal irrelevance (so far) of their own mental states to the corpus they learn to predict. Yet, as we have seen, LLMs might nonetheless develop such capacities as by-products of capacities developed to perform their highly specific linguistic task. This suggests that we should be careful to avoid an overly simplistic application of life-history complexity considerations to AIs.

We are nonetheless excited about the prospects for the life-history approach. An intriguing prospect is that it could be used to theorize, not only about artificial introspection in general, but also about more specific introspective devices. For example, B&V note that social AIs, such as care robots, are likely to need mentalizing capacities, which could then be self-applied to confer introspective abilities. If this is correct, it not only forecasts that social AIs are likely to develop introspection, but also that social AIs will develop introspective devices that fall on the *indirect* side of the direct–indirect dimension of variation we identified. Other-directed mentalizing systems will take behavioural and environmental information as input, and the self-application of such systems would likely involve similar, indirect inputs.

Conversely, we might hypothesize about the types of task that are likely to require *direct* introspective mechanisms. Consider, for example, a robot whose task requires it to have a highly accurate understanding of its own epistemic situation, but which inhabits an environment where illusions and deceptions abound. Such a robot would need to represent its internal states independently of its representations of external states, and would therefore benefit from possession of a direct introspective device. Such a 'Cartesian' robot would still introspect reliably even when its perceptual systems were massively deceived and its representations of external events wholly unreliable. Furthermore, it would introspect confidently even when having low confidence in its beliefs about external events. In his contribution (discussed above), **Fleming** noted the importance of 'introspective robotics' for AI systems operating in novel environments, where forms of metacognition, such as uncertainty monitoring, are needed in order to 'avoid the pitfalls of overconfidence'. The Cartesian robot's needs are an extreme version of this. Fleming also suggests that fine-grained introspective discriminations (beyond low-dimensional measures of uncertainty) might be required in AIs that need to communicate their internal states to others in order to promote epistemic and practical collaboration. Thus a population of cooperative Cartesian robots might require direct and fine-grained introspective devices, of the kind which some (but not all) theorists take human beings to possess.

3.4 Imaginary minds

Pete Mandik brings a speculative, science-fiction perspective to the discussion. He imagines future humans equipped with *Sliders* — brain implants that allow them to call up images of slider controls by means of which they can precisely measure and effortlessly modify their own mental states (Mandik notably mentions moods). These Sliders give their users a very high degree of both introspective self-knowledge and self-determination. Mandik presents the scenario as a cautionary tale; building on the work of science-fiction author R. Scott Bakker, he suggests that the use of Sliders would lead to catastrophe, since the ability to tamper at will with our emotions would undermine the regulatory function of emotion, leading to uncontrollable psychopathy.

Mandik draws two lessons for our project: first, that introspection must be understood in the light of the behaviours it enables, and, second, that if we want to improve our introspective abilities, it is far safer to rely on the slow and effortful contemplative methods we have already devised.

We shall not take a stance on Mandik's prognosis regarding the social impact of Slider enhancements but concentrate instead on what his perspective brings to the study of possible introspective systems, focusing on two points.

First, we welcome the use of science fiction scenarios to explore radically new introspective systems. As we stressed in our target article, an extension of the imagination is needed, and science fiction is an excellent tool for this. Mandik's Sliders, for instance, implement a technological version of what we called *self-applied social perception* — a form of introspection in which a subject *perceives* their own mental states.[8] We noted that the possibility of such a form of introspection was underexplored, and we are glad to see it fleshed out in Mandik's contribution.

However, we note that Mandik's conception of Sliders remains, in some respects, rather conservative: the mental states represented by Sliders (courageous–cowardly, sad–happy, etc.) are folk-psychological ones, and the introspective repertoire afforded by Sliders is not much different from ours. Mandik does envision the possibility that Sliders could decompose 'fractally' into more fine-grained versions, up to a point where they no longer have natural language descriptors. However, this would make their introspective repertoire more precise than ours but not entirely heterogenous. A more radical — and perhaps more interesting —

[8] As Mandik pointed out to us, this could be contested, since we can think of Sliders' users as *seeing* slider controls in their visual field and then merely *inferring* their mental states. But arguably, the inference could become so routinized that the result is more like conceptualized perception ('seeing as'). On this view, an expert Slider user would *see* that they have a certain degree of anger, in the same way that an expert birdwatcher *sees* that there is a tawny owl perched on a branch (rather than inferring it from their perception of shapes and colours).

scenario involves Sliders that latch on to high-level, coarse-grained patterns in our mental lives distinct from those grouped and characterized by our existing introspective repertoires and not represented in folk psychology — for which we might currently have no common word. If these patterns offered significant explanatory and predictive leverage, Sliders that enabled us to identify and control them might be highly useful, and they would constitute introspective devices with repertoires radically different from ours. Of course, it is hard to say what these repertoires might be without substantive theorizing about mental architecture, but again a science fiction perspective could help extend our sense of possibilities.

Second, we have a comment on the first of Mandik's two lessons — that introspection is intimately linked to self-determination. From an evolutionary perspective, Mandik notes, introspection is valuable only to the extent that it facilitates certain adaptive behaviours (typically, kinds of self-determination). We agree (our own definition of introspection ties it closely to the online control of behaviour), but Mandik makes a further point which may be slightly misleading. He suggests that the danger of Sliders lies in their capacity to satisfy our desire for self-knowledge. Such a desire, he suggests, is good only in a context where self-knowledge is hard to get, just as a desire for sugar is healthy only where sugar is scarce. Sliders would be dangerous because they would make self-knowledge abundant, like sugar in modern societies, rendering our self-curiosity unadaptive and even disastrous. We think this mislocates the problem. First, it is debatable that we really have evolved a notable desire for self-knowledge — otherwise, philosophers would not have had to make such *efforts* to follow the Delphic adage 'Know thyself'. (No one ever made efforts to follow the motto 'Eat sweet things'.) Moreover, it is debatable that Sliders would be dangerous because of the self-knowledge they would afford. The danger would lie in the fact that they would enhance *both* self-knowledge *and* self-determination. To see this, imagine 'inert' Sliders that fulfil only the first role — that allow you to visually perceive, say, the exact degree of courage you currently have without allowing you to directly modify it. Such devices would still afford some additional self-control (knowing your exact level of courage would enable you to make better choices about what risks to take), but this would be nothing like the self-determination-at-will whose horrific consequences Mandik describes. In short, the peril of Mandik's Sliders seems to stem almost entirely from the capacity for easy self-determination they offer. Thus, a better evolutionary story might go like this: we evolved strong desires to enter certain mental states (pleasure, euphoria, boldness, guilt-free states, etc.), which were associated with, or proxies for, processes that were fitness-inducing in our ancestral environment. Moreover, there was no need for evolution to bridle our desires for such states, since the risk of achieving excessive levels of them was almost zero given the regulatory capacity of our emotions. Sliders, like the other sorts of wireheading technology mentioned by Mandik, break the reliable connection between mental states and fitness-inducing processes, and undermine the regulatory power of the emotions. Their danger stems from enhanced self-determination rather than from enhanced self-knowledge.

Like Mandik, **Eric Schwitzgebel and Sophie Nelson** (henceforth S&N) use science fiction to investigate the forms introspection might take. Inspired by Ann Leckie's novel *Ancillary Justice*, they imagine an artificial 'ancillary' mind composed of a main computer and two hundred humanoid robots. These components are functionally integrated just enough for it to be indeterminate whether the system is a single, unified mind or a collective of individual minds, and thus indeterminate whether certain representational processes within it constitute introspection within a single mind or metacognitive communication between separate minds. S&N argue that research on possible forms of introspection must capture this in a further dimension of variation, corresponding to whether the process tilts more towards introspection (a mind representing itself to itself) or communication (a mind representing itself to another mind), with a variety of intermediate positions between the extremes. S&N investigate other implications of the ancillary mind scenario that extend beyond our research programme, notably that the number of *persons* present in an ancillary mind situation is also indeterminate. They show that attempts to evade this counterintuitive consequence, invoking various principles (the Body View, the Phase Transition View, and Discrete Phenomenal Realism) are all unsuccessful.

As with Mandik's contribution, we applaud the use of science fiction to speculate about possible introspective systems. However, we have reservations about the suggestion that we should treat introspective–communicative as a further dimension of variation among possible introspective devices. For one thing, by definition, processes that fall clearly on the 'communicative' end are indeed communicative, not introspective, and therefore do not correspond to a way an *introspective* device could function. Second, although S&N show that a process could be indeterminate between introspection and communication, we can also imagine a process that is determinately *both* introspective *and* communicative at the same time. The outputs of a metarepresentational device might be fed both to the central system of the mind in which it is embedded *and* to some other mind (and in both cases be used for online behavioural control). More radically, a person's introspective capacities might be subserved by a communicative process. Suppose that Sam has lost her own introspective capacities, but that her friend Sadie has become expert at inferring Sam's thoughts and feelings from behavioural cues and communicates this information to Sam so rapidly that Sam can use it for online behavioural control. This process counts as introspective by our definition. Yet it is also communicative. Such possibilities speak against treating introspective–communicative as a dimension of variation in introspective devices themselves.

This brings us to another conclusion of the ancillary mind scenario, not explicitly drawn by S&N, but seemingly implied, which could jeopardize our research programme. The ancillary mind case suggests that the classification of a process as introspection rather than metacognitive communication can be indeterminate, depending not only on the nature of the process itself but also on its degree of integration within a wider system. The very same device, with the same internal functioning, could be communicative or introspective (or in-between, or both) depending on the context in which it is embedded. This in turn suggests that introspection and metacognitive communication correspond to less robust natural kinds than the more general process of metacognitive representation, which encompasses both. Hence (the objection goes), when exploring the functioning of minds different from ours, the concept of metacognitive representation should be preferred to that of introspection, and a research programme on possible metacognitive representational systems preferred to one on possible introspective systems.

Whether this conclusion is justified depends on various factors. One is the prevalence and likelihood of indeterminate cases (accepting that the ancillary mind case shows they are possible). Are these cases merely exotic possibilities, or are they likely to abound in natural and artificial environments? There is a case for thinking they will be rare and marginal. It is plausible that both individual *minds* and *collectives of minds* correspond to stable structures, which function as attractors in the space of possible cognitive systems, while intermediate structures are less stable, often evolving into, or being replaced by, one of the two attractors. For a comparison, consider the political domain, where monarchies, oligarchies, and democracies represent stable and widespread governing systems. Conversely, systems with three or four individual powerful rulers are possible but tend to quickly evolve into monarchies — hence their rarity and their rapid dissolution when attempted (as witnessed in Ancient Rome's Triumvirate and Tetrarchy, or Revolutionary France's First Consulate). If it is true that individual minds and collective minds function as attractors, then this warrants independent study of introspection and metacognitive communication. Of course, it might not be true, and research on group minds in nature (e.g., ant colonies) or the future development of artificial ancillary minds could challenge it.[9]

Note, finally, that even if indeterminate minds are stable and likely to be abundant, this would not suffice for the objection to go through. The extremes of a continuum may still be salient for theoretical purposes, even when many cases fall between them. Think for example of political science, where the categories 'liberal' and 'conservative' are useful, even though most subjects are neither typical liberals nor typical conservatives.[10] While matters of theoretical usefulness are difficult to adjudicate without considering particular cases, it is at least plausible that it will often be useful to categorize a process as primarily introspective *or* primarily communicative (or intermediate, or both). rather than prioritizing the general category of metacognitive representation.

4. Lessons and suggestions

In this final section, which can be thought of as an addendum to our target article, we summarize some of the lessons we have learned from the contributors and make some new suggestions arising from the discussion.

4.1 Lessons learned

We have already acknowledged the soundness of many of the points made by the contributors; here we shall focus on ideas for enriching our map of possible introspective systems.

We originally identified three dimensions of variation in the way introspective devices operate: directness, flexibility, and conceptuality. We shall call these *dimensions of operation*. We now think that two modifications to these dimensions may be in order.

First, as we noted earlier, Wu's contribution indicates that we may need to subdivide or specify our existing dimensions of flexibility and directness. The flexibility dimension could be broken down into subdimensions of control (agentive–reflex) and modifiability (modifiable–non-modifiable), the former echoing Renero's account of the *selective route* of introspection, and the latter corresponding to Billon's concept of *tool-flexibility*. Meanwhile,

[9] For example, Luke Roelofs (forthcoming) argues that future technology is likely to create AIs or enhanced humans with 'porous minds', which, like ancillary minds, would engage in processes intermediate between introspection and communication.

[10] Thanks to Sophie Nelson for making this point and providing this example.

the directness dimension could be specified by counting the informational channels mobilized by each device and providing a measure of causal directness for each of them.

Second, the discussions of the measurement and accuracy of introspection by Dołęga, Fleming, and Huebner and Kachru suggest the need to add a new dimension of *performance* to the list of dimensions of operation, comprising such features as accuracy, precision, reliability, speed, and efficiency. In fact, accuracy and precision can already be captured in our framework by reference to the groupings and characterizations permitted by a device's introspective repertoire (see Section 2.2 of our target article). Precision is a matter of how fine-grained a device's groupings and characterizations are, while accuracy (to the extent that it is distinct from reliability) is a matter of whether a device correctly characterizes the states it groups. The other three features are not already captured, however, so we propose to add a performance dimension, broken down into subdimensions of *reliability* (tendency to misfire or malfunction), *speed* (relative to other cognitive operations), and *efficiency* (how effectively a device mobilizes energetic and computational resources to do its work).

Another upshot of the symposium is that we may need to recognize a further set of dimensions of variation, corresponding to how introspective devices are deployed within the wider cognitive system. We shall call these *dimensions of deployment*, and we shall mention three of them. Note that we define dimensions of deployment for individual introspective *devices*, rather than for introspective *systems* (which may be composed of multiple devices).[11]

First, Carruthers and Masciari have persuaded us of the possibility (and perhaps actuality) of subpersonal introspective processes, and thus of the option of recognizing a further dimension of *cognitive access*, corresponding to how widely accessible the outputs and activity of an introspective device are within the cognitive system. This might have two subdimensions, one specifying whether the *contents* represented by the device are available at a personal level, and the other whether the *activity* of the device is so available (the latter corresponding to the dimension of *opacity* mentioned by Dołęga).

Second, Schwitzgebel and Nelson's contribution raised the possibility of adding an introspective–communicative dimension of deployment, reflecting whether a metarepresentational device provides information to the system in which it is embedded or to

[11] We do this because an introspective system may be highly disunified, composed of many independent devices deployed in different ways, and we need a taxonomy that captures this. Of course, this device-focused perspective may be less useful when considering highly unified systems, in which a small number of devices are tightly coordinated and deployed as a unit. The devices composing such systems will all have similar deployment characteristics, and it might even be useful to treat the systems as single complex devices.

another one. In some contexts this may be useful, though the dimension will not be universally applicable, since, as we saw, a metarepresentational device could simultaneously play both introspective and communicative roles.

Third, it may be helpful to map variation in *function*, since, as noted by Huebner and Kachru and highlighted by Browning and Veit's application of life-history considerations, introspective devices may play diverse roles. In our target article we defined introspection as providing information that can be used for online behavioural control, but we did not claim that introspection only serves that function, and of course online control itself can take a variety of forms. It may therefore be worth detailing some of the specific functions an introspective device might perform. We could start with a single–multiple functions dimension, supplemented with a range of specific functions, including self-regulation, self-prediction, epistemic evaluation, cooperation, maintenance of a sense of self, and therapeutic self-shaping (as discussed by Huebner and Kachru). This list could doubtless be extended.

Table 1 recapitulates the updated list of dimensions of introspective devices we propose. The starred categories are ones mentioned in our original paper; the rest are additions and refinements inspired by the symposium contributions. Question marks indicate that we do not currently take a stance on whether and what subdimensions/specifications might be useful.

[Insert Table 1 about here]

In addition to these dimensions of variation in *devices*, our target article also identified a higher-level dimension of variation in introspective *systems*, specifying how coordinated a set of introspective devices is. We called this dimension *unity*. Billon's reflections on the potentially 'plurimodal' nature of introspection suggest that it might be useful to break this down into two components (something we hinted at in our article): a measure of the *number* of distinct introspective devices, and a measure of the *coordination* of these devices (which will have to factor in the similarity of their repertoires). This is summarized in Table 2.

[Insert Table 2 about here]

Of course, as we noted when discussing Spener's and Wu's contributions, the choice of which dimensions to centre in any given case will depend on theoretical context. Depending on one's goals and interests and on the exact phenomenon one is studying, different maps, with different breakdowns of the dimensions, will be more or less enlightening.

4.2 New suggestions

In this section we shall briefly mention some themes that were not substantively treated by the contributors to the symposium, but which we think deserve further exploration if research on possible introspective systems is to be fully developed. One thing we call for is more exploration of intercultural and interlinguistic variation in introspection among humans (possibly, but not necessarily, linked to variation in theory of mind). We would also like to see more detailed exploration of the forms introspection might take in embodied AI systems designed to perform various tasks or embedded in various environments. Several contributions, including those by Fleming and Browning and Veit, hinted that we should expect some AI systems to need and develop introspective capacities of various kinds. Case studies focusing on introspection in social AIs, or robots operating in high-uncertainty situations, would be a useful complement to Long's detailed approach to LLM introspection.

One topic that came up frequently in our responses to contributors is that of introspective repertoires. While a delineation of possible introspective repertoires occupied a substantial portion of our original paper, few contributors commented on introspective repertoires, and even fewer seriously explored possible variations in them. Contributions which directly discussed the content of introspective representations (Wu, Dołęga) left the issue of possible variation in introspective repertoires largely uninterrogated, and even imaginative and speculative pieces (Mandik, Schwitzgebel and Nelson) remained relatively conservative in their accounts of the introspective repertoires possessed by the imaginary minds described. We were also somewhat disappointed that contributors who discussed the accuracy and precision of introspection (Dołęga, Fleming) didn't explore the utility of thinking of these matters in terms of the discriminations and characterizations made by different introspective repertoires.

It may be that the relative paucity of discussions of introspective repertoires signals that there is not much to say on the topic. Alternatively, it might show that no satisfying method has yet been found to explore variation of this kind, and that creative thinking is needed here. At any rate, we suspect that progress could be made in this direction, and we hope that researchers will take up the challenge in the years to come.

Conclusion

This symposium has been a stimulating exercise for us. We have learned much from reflecting on the contributions, and though some contributors have expressed doubts about our project, the overall effect has been encouraging. We are reassured that a systematic investigation of possible forms of introspection can broaden our sense of theoretical possibilities, shedding new light on familiar forms of introspection and helping us to recognize and understand unfamiliar forms. We hope that the exercise has had a similarly positive effect on the readers of this special issue and stimulated them to contribute to research on what forms introspective systems could take.

References

Baird, B., Mrazek, M. D., Phillips, D. T. & Schooler, J. W. (2014). Domain-specific enhancement of metacognitive ability following meditation training. *Journal of Experimental Psychology: General*, *143*(5), 1972–1979.

Beck, B., Peña-Vivas, V., Fleming, S., & Haggard, P. (2019). Metacognition across sensory modalities: Vision, warmth, and nociceptive pain. *Cognition*, *186*, 32–41.

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, *23*(11–12), 11–39.

Kammerer, F. (2021). The illusion of conscious experience. *Synthese*, *198*, 845-866.

Katz, L. D. (2016). Pleasure. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), URL = <https://plato.stanford.edu/archives/win2016/entries/pleasure/>.

Marcus, G. & Davis, E. (2020). GPT-3, bloviator: OpenAI’s language generator has no idea what it’s talking about. *MIT Technology Review*. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

Morales, J. (forthcoming). Introspection is signal detection. *British Journal for the Philosophy of Science*, https://doi.org/10.1086/715184.

Pasquali, A., Timmermans, B. & Cleeremans, A. (2010). Know thyself: Metacognitive networks and measures of consciousness. *Cognition*, *117*(2), 182–190.

Péron, F. (2012). Language-trained animals: A window to the ‘black box’. *International Journal of Intelligence Science*, *2*(4), 149–159.

Roelofs, L. (forthcoming). Porous minds: When does communication become introspection? In R. Sterken & H. Cappelen (eds.), *Communicating with AI: Philosophical Perspectives*. Oxford University Press.

Rossi, A. P. & Ades, C. (2008). A dog at the keyboard: Using arbitrary signs to communicate requests. *Animal Cognition*, *11*(2), 329–338.

Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science*, *1*, 253–329.

Treffert, D. A. (2009). The savant syndrome: An extraordinary condition. A synopsis: past, present, future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1522), 1351–1357.

Tables

Type of dimension	*Dimension*	*Possible subdimensions or specifications*
Dimensions of operation	Conceptual–non-conceptual*	?
	Flexible–inflexible*	Agentic–reflex
		Modifiable–non-modifiable
	Direct–indirect*	Specification via a count of the informational channels used by the device, with a measure of directness for each
	High–low performance	Reliable–unreliable
		Fast–slow
		Efficient–inefficient
Dimensions of deployment	Accessible–inaccessible	Personal–subpersonal (how accessible introspective contents are)
		Opaque–transparent (how accessible introspective activity is)
	Introspective–communicative	?
	Single function–multiple function	Self-regulation: Often–never
		Self-prediction: Often–never
		Epistemic evaluation: Often–never
		Cooperation: Often–never
		Maintenance of a sense of self: Often–never
		Therapeutic self-shaping: Often–never
		Etc.

Table 1: Dimensions of variation in introspective devices

Type of dimension	*Dimension*	*Possible subdimensions or specifications*
Higher-level dimensions	Unified-disunified*	One–many devices
		Coordinated–uncoordinated (if more than one device)

Table 2: A dimension of variation in introspective systems