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

What forms could introspective systems take? A research programme[1]

Authors

François Kammerer (Ruhr-Universität Bochum)

Keith Frankish (University of Sheffield)

Abstract

We propose a new approach to the study of introspection. Instead of asking what form introspection actually takes in humans or other animals, we ask what forms it *could* take, in natural or artificial minds. What are the dimensions along which forms of introspection could vary? This is a relatively unexplored question, but it is one that has the potential to open new avenues of study and reveal new connections between existing ones. It may, for example, focus attention on possible forms of introspection radically different from the human one and help to integrate competing theories of human introspection in a non-adversarial manner. We

[1] The authors thank David Chalmers, Antonio Chella, Eileen Frankish, Elizabeth Irvine, Bryce Huebner, Russell Hurlburt, and Eric Schwitzgebel, as well as two anonymous reviewers for the *Journal of Consciousness Studies* and audiences at Bochum, Fribourg, Liège, and Rethymno.

introduce and motivate the project, provide a preliminary mapping of the space of possible forms of introspection, and sketch a programme for interdisciplinary research on the topic.

Introduction

There is much debate among philosophers and scientists about what introspection *is*. What exactly do we mean by 'introspection'? How does introspection work? What mechanisms does it use? How does it compare with perception? Is it distinct from self-interpretation and mindreading, and, if so, how? What kind of knowledge does it provide, and how reliable is it? What does it tell us about the nature of the mental states it tracks? Does 'introspection' pick out a single process, or is it an umbrella term for a variety of processes?

Here is a question that has attracted less attention, and which we intend to explore here: *What forms could introspective systems take?* That is, regardless of the nature of human introspection, what are the possible ways in which a cognitive system could introspectively represent its own current mental states? Theories of human introspection can of course help us answer this broader question. However, it remains a fundamentally different research question, and many self-representational processes that are not plausibly possessed by humans could nevertheless constitute genuine ways in which a cognitive system could learn about its own mental states.

Our goal is not to give a full answer but to pave the way for a systematic study of the issue by defining the space of possible forms of introspection and drawing up an agenda for a research programme on the topic. By considering what introspection could be, we aim to construct a framework for representing possible forms of introspection. Researchers may of course use this framework to locate and compare competing views of human introspection.

However, we anticipate that it will more often be used to think about the introspective processes of other creatures, including non-human animals, enhanced humans, artificial intelligences, and aliens.

First, we define our research question, explain its motivation, and outline our approach. (§1). Second, we map the space of possible forms of introspection by detailing some of the main dimensions along which introspective processes might vary (§2). Third, we address objections to the view that introspective processes might vary in fundamental ways (§3). Finally, we suggest areas for future research, both empirical and theoretical, and sketch a research programme on possible forms of introspection (§4).

1. Defining and motivating the project

1.1 Defining introspection

We shall start with the following working definition of introspection:

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

There are three conditions in this definition. The first concerns the *target* of introspection: introspective processes are directed at the system's current mental states. The second concerns the *nature* of introspection: introspective processes are representational. The third concerns the *function* of introspection: introspective processes enable a cognitive system to use information about its own mental states for online behavioural control.

Some notes on these conditions. We use 'representational' in an inclusive sense that is neutral between different views of mental representation (although our approach is naturalistic). We use 'mental states' in a similarly inclusive sense. We take paradigm examples of mental states to be folk-psychological states (beliefs, desires, intentions, perceptions, sensations, emotions, moods, and so on), but we also include variants and analogues of these states posited by psychologists and roboticists, as well as hypothetical variants and analogues that might be attributed to AIs and aliens. Moreover, we intend our account to be compatible with a wide range of views about the precise nature of the various types of mental states and how they are realized in particular cognitive systems. By *current* mental states we mean both transient states (such as perceptions, emotions, and 'occurrent' beliefs and desires) and persisting states (such as long-term memories and 'standing state' beliefs and desires) that are not currently active.

The 'target' condition should be understood *de re*: when we say that introspective processes represent mental states, we mean that they represent states *that happen to be* mental states, but not necessarily that they represent them *as being* mental states. (A *de dicto* understanding would unnecessarily limit the possible forms of introspection; see Section 2.2.)

By 'online behavioural control' we mean the regulation of current behaviour. The system might use introspective information to guide social behaviour, sharing or withholding information about its mental states in order to produce certain effects in others. Or it might use it to exert higher-level self-control, regulating how its mental states affect its behaviour. For example, if a person becomes introspectively aware that they are strongly tempted to do something they know they would later regret, such as overeating or starting an argument, they may act to put themselves out of the reach of temptation. We specify that the information should be available for *online* use in order to capture the idea that introspection is a mental faculty that can feed directly into the control of behaviour. However, we do not mean that the information it supplies is available *only* for online use. Self-knowledge produced by introspection may be stored and used to guide later behaviour. (Knowing from past introspective episodes that one has a fondness for unhealthy food, one may adopt of a policy of avoiding places where it is served.) Neither do we mean that introspective information *must be* used for behavioural control: some (that obtained in meditation, perhaps) may never be. Our claim is merely that is *available* for such use. Finally, we assume that introspective information will be globally accessible within the cognitive system (at a ‘personal’ level), and that it will usually (though perhaps not always) be used in practical and theoretical reasoning. That is, we assume that introspection typically generates metacognitive *beliefs* (as opposed to subdoxastic states) or at least metacognitive states, such as metacognitive feelings, that are directly available for the formation of metacognitive beliefs (as opposed to being completely inaccessible subpersonal states).

This working definition is purposely intended to be a liberal one. It includes all processes standardly accepted as introspective ones, such as those described by inner sense theorists (Armstrong, 1980) and transparency theorists (Byrne, 2018). It might be taken to exclude the relation described by *acquaintance* theorists (Chalmers, 2003; Gertler, 2012), since acquaintance is typically understood as a primitive, non-representational epistemic relation between subjects and their phenomenal properties. However, if we adopt a liberal enough sense of ‘representation’, synonymous with ‘presentation’ or ‘grasp’, we might include acquaintance as a limit case. Moreover, since our definition does not require introspection to be distinctively first-personal or non-inferential, it also includes processes of self-representation that many would not regard as genuinely introspective, including ones involving the self-application of a naïve theory of mind or general principles of self-interpretation (Carruthers, 2011; Dennett, 2017; Frith & Happé, 1999; Gopnik, 1993; Graziano, 2013).

Precisely because this definition is so liberal, it may seem unsatisfying to theorists who think that introspection should be characterized in a richer way, by reference to the first-person perspective, the phenomenal mind, the privacy of the mental, the authority of the subject, and so on. While we acknowledge that our definition is more liberal than most traditional ones, we ask the reader to bear with us and consider the questions that this liberal definition allows to ask. A richer conception of introspection rooted in self-reflection might incorporate human-centric biases and assumptions, making it a poor tool for thinking about introspection in other creatures or even about human introspection itself. (As illusionists about phenomenal consciousness, we take seriously the idea that we may be deeply mistaken about our own minds.) By starting with a minimal conception of introspection, we hope to avoid these risks. We shall say more about the issues raised by our definition in Section 3.

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

In this liberal sense, it is undeniable that humans introspect, and it is not implausible that at least some non-human animals do (see Section 4). However, there is still much disagreement regarding the exact nature of human introspection. Does introspection employ the same process used to represent the mental states of others, or does it employ a distinct one? Is introspection perception-like, or does it involve the application of conceptual representations? How accurate is introspection? How unified is it? (For an overview, see Schwitzgebel, 2014.)

We propose to turn aside from these debates, however. Instead of trying to zero in on the actual introspective processes found in humans, we propose to widen the focus and explore the space of *possible* introspective processes.

1.2 Motivating the project

What are the possible ways in which a cognitive system could represent its own current mental states in a manner that allows for online use of the information? There are several reasons for considering this question. We shall mention four, in increasing order of importance.

First, the question is interesting in itself, and since it has never been asked in a systematic manner, it is worth dedicating at least *some* attention to it; the inquiry might be productive of unexpected insights.

Second, the question invites us to take a new look at existing theories of human introspection. Most of these contradict each other, so they cannot all be true, but they may all still describe possible ways in which a mind could introspect. Asking what introspection *could be* may lead us to build an encompassing framework in which these theories can be located and compared, perhaps highlighting unobvious similarities and differences between them. In doing so, the project should also help to promote interdisciplinary engagement and collaboration. Mapping the space of possible introspection will provide a conceptual and terminological framework for comparing and contrasting models of introspection developed in different fields and for different purposes.

Third, the project should extend our sense of possibilities. Exploring the space of possible introspection will direct our attention to regions of the space that have so far been neglected. It is particularly important to do this for introspection. Numerous examples make it clear that other forms of perception are possible, differing from those we are familiar with (snakes see infrared, bats perceive by echolocation, pigeons and sharks perceive magnetic fields, etc.). But examples of alternative forms of introspection are far less obvious, and we might assume that the familiar forms, of which we have some intuitive understanding, are the only ones possible. Explicit reflection on other forms introspection could take, or could have taken, employing a liberal notion of introspection unconstrained by assumptions about how we know our own minds, should serve as a corrective, helping us to determine which features of our own introspective processes are necessary and which contingent. This might in turn teach us much about human introspection. For example, it should help us to understand how our introspective mechanisms evolved. Were they the only way to accomplish the relevant functions or could quite different mechanisms have done the same job? If the latter, why were these mechanisms selected for rather than any of the alternatives?

Fourth, and most importantly, the enquiry should help us think about introspection in creatures different from neurotypical humans, including other terrestrial animals and beings we may create or encounter in the future, such as artificial intelligences, enhanced human intelligences, and alien intelligences. We share this planet with creatures very different from ourselves, whose mental capacities we have traditionally underestimated. If we are to understand and appreciate the diversity and complexity of terrestrial minds (including neurodivergent human minds), we shall need to adopt a far less anthropocentric perspective and accustom ourselves to imagining forms of mentality very different from the neurotypical human one. By encouraging researchers to speculate about possible forms of introspection, our project should contribute to this wider task.

At the same time, the project should also help prepare us to encounter new forms of mental diversity. In the decades to come, we shall live among increasingly sophisticated artificial and enhanced minds, and it would be wise to ask in advance how such minds could think, including how they could think about themselves and their own mental states. To explore the space of possible introspection is to explore the ways in which other minds could represent themselves. We see our project as analogous to Bartlett and Wong's project of theorizing about what they call 'lyfe', understood as including 'life as we don't know it', something wider than 'life' – life on Earth, as we do know it (Bartlett & Wong, 2020). Bartlett and Wong argue that defining and theorizing lyfe is an important conceptual preliminary to astrobiological research, and defining and exploring the space of possible introspection should serve a similar role with respect to understanding non-human minds.

1.3 The minimal mind approach

Before we can formulate a research programme to explore the space of possible introspection, we need to provide a preliminary map of the space. We shall adopt an engineering perspective, considering a *minimal mind* and asking how we might equip it with introspective capacities. By a 'minimal mind', we mean a cognitive system which (1) is equipped with sensors and effectors and inhabits a physical and social environment populated with other similar minds, (2) has first-order mental states (say, perceptions, beliefs, and desires), but (3) lacks introspection as we have defined it. We can then ask how we might equip this minimal mind with introspective abilities. From this engineering perspective we should be able to map the space of possibilities in a way that is not too tightly constrained by our intuitive understanding of our ordinary, human capacity for introspection.

The next section is devoted to mapping the space of possibilities in three ways, focusing on introspective devices (2.1), introspective repertoires (2.2), and the unity of introspection (2.3).

2. Exploring the space of possible introspection

2.1 Possible introspective devices

Equipping our minimal mind with introspection means giving it a new representational device of some kind, and we shall have to make choices about the features of this device, including its inputs, its internal functioning, and its outputs. We shall not attempt an exhaustive catalogue of these choices, but here are a few dimensions along which choices would have to be made.

(1) *Direct–indirect*. Starting with the inputs, we shall have to decide how *close* and *direct* the informational relation will be between the introspective states and the mental states they represent. At one extreme, there are direct representational systems, where the tokening of the represented first-order states proximally cause, or even constitute, the tokening of the representing introspective states, and the informational dependence of the latter on the former is close and direct. Think, for example, of how a mercury thermometer represents temperature. At the other extreme, there are highly indirect representational systems, where the informational dependence between representation and represented state is highly distal and mediated. Think about the steps involved in the process by which a few pixels on the screen of a smartphone running a weather app represent the outside temperature. Of course, there are many possibilities between these extremes. (For a sketch of a possible highly indirect form of introspection, see the discussion of self-applied social perception below.)

(2) *Nonconceptual–conceptual*. Turning to the output, we shall have to decide whether the representational device will generate representational states with a format that is *conceptual*, akin to beliefs or propositionally structured perceptions, or *nonconceptual*, akin to sensations. We don’t see this as a binary distinction, however, but as a matter of degree, a dimension. At

one extreme, we have the most conceptual kind of output: say, complex belief states that possess digital content (Dretske, 1981), have systematically recombinable components (Fodor & Pylyshyn, 1988), and are highly inferentially integrated with the rest of the system's beliefs. At the other extreme, we have the most nonconceptual output: say, sensory states that have purely analog content, lack recombinable components, and have no antecedent inferential integration. Many other possibilities lie in between. For example, sensory states which have analog content but are strongly disposed to activate conceptualized beliefs would be more conceptual than the most nonconceptual states. And belief states which have digital content and recombinable components but are inferentially isolated from other beliefs would be more conceptual still, though not maximally so.[2]

(3) *Inflexible–flexible*. Third, we shall have to make choices about the internal functioning of the introspective device. These could be made along many dimensions. One is *flexibility*. At one extreme, we can imagine a highly flexible device, whose internal functioning can be directly and deeply modified by the system of which it is a part. This would be the case, for example, if central aspects of its functioning were under intentional control — responsive to the beliefs, desires, and intentions of the mind of which it is a part. At the other extreme, we can imagine a very inflexible introspective device, whose internal functioning cannot be directly affected by the wider system. (Of course, it might still be *indirectly* affected by it; voluntary self-modification is always possible.) We can illustrate this dimension with examples from outside the introspective domain. The mechanisms by which we represent

[2] As a reviewer pointed out, we could isolate three dimensions here: conceptuality per se, recombinability of components, and inferential integration. For illustration purposes, however, we choose to integrate them into a single thickly conceived *conceptual/non-conceptual* dimension.

social norms, for example, are very flexible. We can easily learn new sets of norms, appropriate to different cultures, and switch between them. On the other hand, the mechanisms by which we visually represent the size and colour of objects are highly inflexible, as illustrated by the fact that many visual illusions are cognitively impenetrable.

An introspective device can be more or less flexible, depending on how much of its functioning can be modified, as well as on the ease of such modifications. For example, a cognitive system might be able to control *when* introspection occurs and *where* it is directed (say, whether to beliefs or to perceptions) but unable to control *how* it operates (what processes it uses or what format its outputs take).

There are two things to note here. First, these three dimensions are not the only ones along which introspective mechanisms could vary. For example, they might also vary in their accuracy, speed, and resource consumption, and in how their outputs are used by the wider system. Second, it is tempting to speculate about where an introspective device is likely to lie along these various dimensions. At this stage, however, we are not concerned with this. We simply note that these are dimensions along which *representational devices* can vary, and, therefore, along which possible *introspective devices* can too.

We can now use these three dimensions to create a diagram of *possible introspective devices* (PIDs).

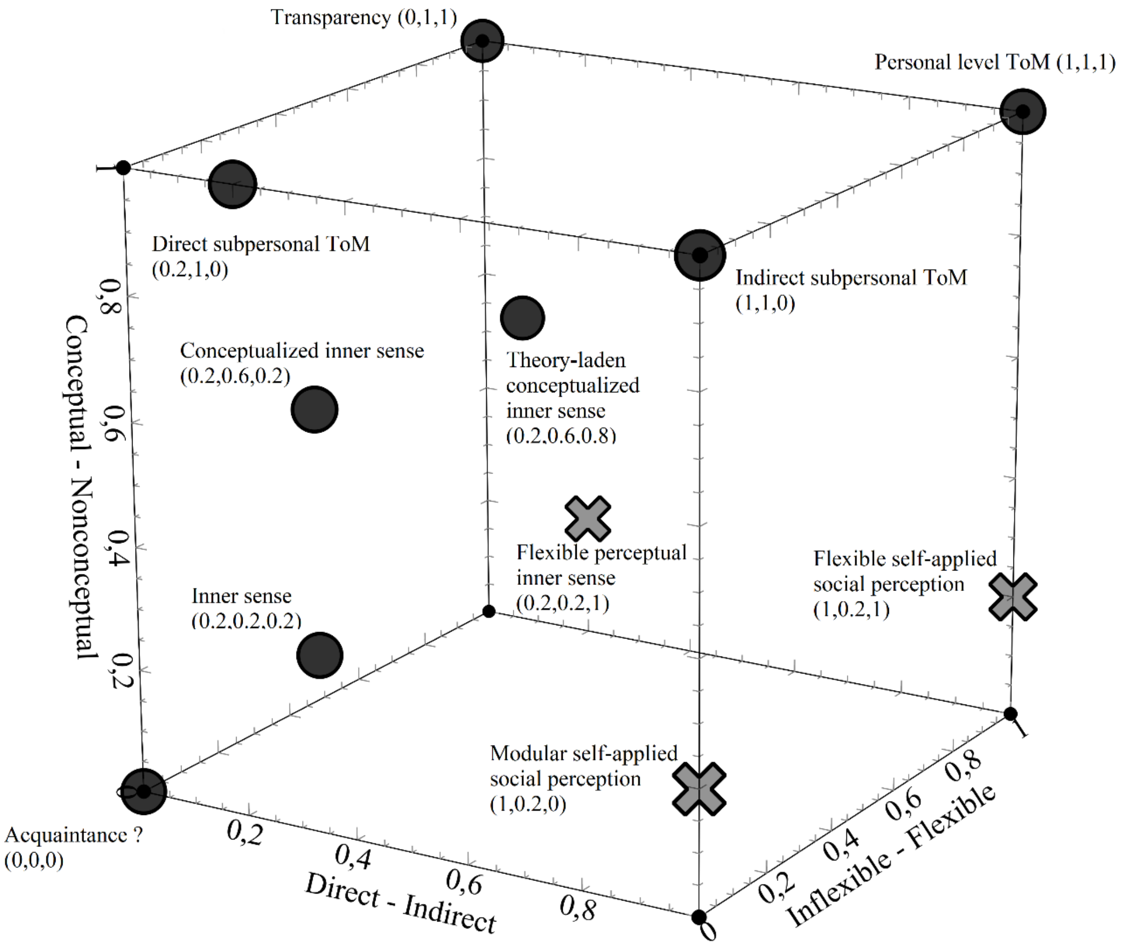

Figure 1: Possible introspective devices:

Existing theories of human introspection can be treated as describing possible introspective devices and marked on the diagram. We have used circles to mark locations for the following views:

- *Inner-sense views*, according to which introspection involves a quasi-perceptual mechanism which directly monitors the system’s first-order mental states (Armstrong, 1980).
- *Conceptualized inner-sense views*, according to which the representations generated by the self-monitoring mechanism are conceptualized (Nichols & Stich, 2003). This category

could also include views that see introspection as involving the direct application of phenomenal concepts by the detection mechanism (Levin, 2007).

- *Theory-laden conceptualized inner sense views*, according to which the representations generated by the self-monitoring mechanism are conceptualized in terms of learned psychological or neuroscientific theory (Churchland, 1985).
- *Transparency views*, on which introspective beliefs are formed via the application of transparency inference rules, such as ‘If P, then believe that you believe that P’ (Byrne, 2018). (This process is said to be *transparent* because beliefs about the system’s mental states are formed by focusing outwards on the world rather than inward on the mental states themselves.)
- *Personal-level theory of mind views*, according to which introspection consists in the personal-level self-application of a flexible theory of mind or flexible principles of self-interpretation (e.g., Dennett, 2017; Gopnik, 1993).
- *Subpersonal theory of mind* views, according to which introspection involves the application of a theory of mind by a relatively inflexible *theory of mind module* (ToMM) (e.g., Carruthers, 2011; Frith & Happé, 1999; Graziano, 2013). This form of introspection can be direct or indirect, depending on what kind of input the module takes. Direct forms have internal access to information about the system’s mental states, whereas indirect forms use information about the system’s behaviour to infer what its mental states are.
- *Acquaintance* views (marked with a question mark), according to which introspection is an unmediated, primitive relation (usually restricted to phenomenal introspection) (e.g., Chalmers, 2003; Gertler, 2012). As noted earlier, these should be thought of as a limit case of possible introspective devices.

There is room for debate regarding the exact location of each view, but one function of the diagram is to help us explore different interpretations of the views, each corresponding to a different PID. Note also that we could use lines, surfaces or volumes to represent *families* of devices in the diagram.

An important heuristic function of this diagram is to direct attention to regions of the space of PIDs that are currently unoccupied by models of human introspection. The crosses indicate three such locations. Two correspond to forms of what we call *self-applied social perception*. ‘Social perception’ is the name for a hypothetical process which represents social features, including psychological ones, in perception itself. The idea is that people can literally *see* that a friend is angry, *hear* that she is sad, and so on (Spaulding, 2015). Now, we can imagine PIDs which involve self-applying such capacities: we would literally see, hear, smell, taste, or feel our own mental states, or, perhaps, feel them in some new sensory modality. (Such a modality would still be distinct from an inner sense mechanism, since the informational relation would be radically indirect.) As a toy example, imagine that we were unable to detect our emotions directly, but that a subpersonal system continually monitored our behaviour and social interactions for signs of emotion and caused us to undergo distinctive colour experiences corresponding to the patterns detected. Our raised voice, our rapid hand gestures, and the frightened reactions of others would cause us to literally see red, alerting us to the fact that we were angry. Similarly, signs of sadness would make us see blue, signs of bitterness yellow, and so on. This description suggests an inflexible mechanism, but we can imagine more flexible versions, in which past experience can modify colour associations and training and attention can produce novel colour sensations, corresponding to new or more fine-grained emotions.

These social perception PIDs would draw on indirect bodily and environmental data and deliver their output in the form of nonconceptual, perceptual representations. Two crosses on the diagram correspond to such PIDs, one to a flexible version, the other to an inflexible, modular version. The third cross marks another unoccupied region, corresponding to a radically flexible kind of inner sense. Think of a mechanism which directly monitors one's mental states and produces nonconceptual representations of them but does so in a flexible way, so that learning, training, and focusing of attention change which representations are activated and how. One could then successively latch on to very different patterns in one's mental life, though the representations involved would remain nonconceptual. (This contrasts with a theory-laden inner sense mechanism, whose representations are permeated by theoretical beliefs.) The kind of introspection afforded by some forms of meditation may come close to this (Dunne et al., 2019), although it is debatable that it offers the right degree of flexibility.

This is just a sketch. More work is needed to see if these unoccupied regions of the space of PIDs represent interesting possibilities (for more on what makes a possibility interesting, see Section 4). Moreover, the number of unoccupied regions would increase if we looked at the diagram with a finer grain of analysis — looking, not at introspection in general, but at the introspection of specific types of mental state. Indeed, some existing theories of introspection were proposed primarily for specific types of state (e.g., self-applied theory for beliefs and desires but not for sensations). Finally, exploring other dimensions along which PIDs could vary would lead to the construction of other diagrams, highlighting other unoccupied regions. Our diagram is merely a first, tentative step, and we encourage such explorations.

2.2 Possible introspective repertoires

Another important way in which PIDs might vary concerns the introspective representations themselves. We said earlier that the outputs of PIDs could be more or less conceptual. We represented this as a dimension of our diagram. However, this variation concerned only the *format* of the introspective representations, not their *content*. We still need to decide what these representations will represent. By definition, they will represent *mental* states (understood *de re*). However, there are possibilities for variation here. We shall highlight two.

First, introspective devices may make different *discriminations* among the mental states they target, tracking different types of state and grouping them in different ways. One device might distinguish emotions from moods and treat them as different mental types, whereas another might group them together. Note that discriminating mental states isn't the same as *conceptualizing* them. An introspective device might distinguish two types of mental state without characterizing them in any substantive way; it might simply represent them as *this type* and *that type*. It need not even characterize them as *mental states*. (Remember that the 'target' condition on introspective representations is to be read in a *de re* manner). However, introspective devices might also be equipped with mental-state concepts, enabling them to generate outputs which characterize the mental states they detect.

Second, different introspective devices might *characterize* the mental state types they distinguish in different ways, using different conceptual schemes. For example, two devices might both distinguish moods from other mental states but characterize them differently — one as moods, the other as (say) a species of perception. Note that we are not assuming that introspective devices will always characterize states *correctly*. There are possible introspective devices that radically mischaracterize the states they detect, and some of these mischaracterizations might even be adaptive.

We can now add more detail to our map of possible introspections. Let us say that the set of discriminations and characterizations that an introspective device can make constitutes its *introspective repertoire*. Then every point within our three-dimensional space of introspective devices can be associated with a further space of *possible introspective repertoires* (PIRs).

Where a given device stands in the space of PIDs imposes constraints on its associated repertoire. If a device lies towards the non-conceptual end of the conceptualization axis, it will be limited in the kind of characterizations its repertoire could provide. Moreover, only devices that fall at the lower end of the flexibility axis will have a fixed repertoire. Indeed, a key way in which an introspective device can be flexible is by having a flexible repertoire.

How do we map the space of PIRs? It won't be easy, since there will be many (maybe infinitely many) possible repertoires, and we may currently have no concepts for many of them, especially those very different from our own. We can make a start, however, by listing putative features of *our own* introspective repertoire: groupings and characterizations that are taken to reflect the way human adults introspect. (Whether or not humans really do introspect in this way is irrelevant; we are using the features simply as a preliminary way of dividing up the space of possibilities.) Here are some such features.

(1) *Direction of fit*. Introspection might group mental states by their direction of fit (Anscombe, 1957), distinguishing those that represent how things are (with *mind-to-world* direction of fit) from those that represent how things are to be made to be (with *world-to-mind* direction of fit). It might also *characterize* the states so grouped as having those directions of fit.

(2) *Perceptual vs cognitive*. Introspection might distinguish perceptual states (sensations, perceptions, perhaps also emotions and feelings) from cognitive states (abstract thinking,

believing, supposing, etc.).[3] Again, this grouping might be accompanied by a characterization of such states as perceptual or cognitive.

(3) *Intentionality.* Introspection might group together those mental states that are intentional (that are *about* something, have a *content*) and thus have correctness conditions. If all mental states are intentional (as some have argued; e.g., Brentano, 1874/2015), then this grouping will coincide with the set of all the mental states detected. If some mental states are intentional and others not, then there will be two corresponding introspective groupings. Moreover, introspection could *characterize*, correctly or incorrectly, some or all of the states it represents as intentional. It is often supposed that human introspection represents perceptual states, imaginative states, and propositional attitudes as intentional, but there is debate about whether it represents emotions and sensations as intentional too.

(4) *Relationality*. Introspection might group together those mental states that are *relational* — that is, consisting partly or wholly in a relation to something. This is a variation on the previous feature, since intentionality is itself a relation (or at least a ‘quasi-relation’[4]). However, intentionality is a specific relation, which can be held to non-existent things and which grounds correctness conditions, and there are other ways in which mental states could be relational. Again, introspection could also *characterize* some mental states as relational (but not necessarily intentional). For example, it is arguable that human introspection represents states of *noticing* as relational (one notices *something*) but not exactly intentional (one cannot notice something that is not there). Various other relations could also be used to

[3] For two views on what grounds the perception/cognition divide, see Beck, 2018; Kriegel, 2019.

[4] See the first section of Brentano’s ‘Appendix to the classification of mental phenomena’ (Brentano, 1874/2015).

discriminate and characterize mental states, including causal relations and resemblance relations.

(5) *Phenomenality*. It is often claimed that at least some mental states possess phenomenality or phenomenal character — a distinctive 'subjective feel' which makes it 'like something' to be in them. If so, then introspection could group such mental states together. And if some mental states do not possess phenomenality, then introspection could make two groupings and a corresponding distinction. The authors believe that phenomenality is illusory, and thus that it cannot really be the basis for an introspective grouping (e.g., Frankish, 2016; Kammerer, 2021). However, an introspective device might still *characterize* some mental states as phenomenal, though, if we are right, it would be doing so on the basis of some other 'quasi-phenomenal' feature (Frankish, 2016). It is usually supposed that human introspection represents at least sensory, perceptual, and emotional states as phenomenal, although some argue that it also represents cognitive states as phenomenal (e.g., Bayne & Montague, 2011).

Now, we can use these groupings and characterizations to classify possible introspective repertoires. Does a repertoire discriminate states with different directions of fit? Does it group together intentional states? Does it characterize some introspected states as phenomenal? And so on. Table 1 below uses these features to classify some real and imaginary introspective repertoires.

We could refine our classification by asking more specific questions. If an introspective repertoire distinguishes perceptual states and cognitive states, does it make further distinctions within each group — say, between visual perceptions and auditory ones or between beliefs and suppositions? And does it *characterize* these states as such? Or, if an introspective repertoire characterizes some mental states as phenomenal, does it make distinctions among these states and characterize them accordingly? An obvious candidate for such a distinction would be one based on *valence* (between pain, pleasure, etc.).

So far, we have focused on features we attribute to our own introspective repertoire. But we can extend the method to introspective repertoires different from ours. As an illustration, here are some features that are plausibly not used as the basis for groupings and characterizations by human introspection, but which could be used by other introspective devices.

First, consider *energy cost*. Being in a mental state incurs some energy cost for the system, and different states have different costs. Doing mental arithmetic, attending to a rapidly changing scene, or holding detailed information in memory are likely to have a high energy cost compared to, say, daydreaming or listening to relaxing music. We do not seem able to introspect this feature of our mental states, but we can imagine a PID which can, and which is equipped with a PIR suitable for making the corresponding groupings and characterizations. (Think of how easily a personal computer can monitor how much Random Access Memory is being used.)

As a second example, think of the *genealogical* properties of standing states. We humans have all kinds of standing beliefs, desires, and hopes, but we cannot introspect the timing and circumstances of their generation and modification. Nor do we have corresponding introspective groupings and characterizations (say, *beliefs formed before 12* vs *beliefs formed after 12*). However, we can imagine a system endowed with an introspective device that does just that (again, think of how easily your personal computer keeps track of the timing of modifications to its files).

We can also imagine creatures whose introspective repertoires differ radically from our own. Here are sketches of three such creatures.

A *phenomenalist* possesses a single introspective device, with a repertoire exclusively composed of phenomenal concepts. Its introspective device represents a variety of internal states but characterizes them all as phenomenal states. (Whether or not these states really are phenomenal ones is beside the point.) Each type of state is characterized as a distinct primitive type of phenomenal state, as distinct from the others as phenomenal red and phenomenal green are for us. Thus, when a phenomenalist introspects, it forms beliefs of the form, 'I am currently in state X / state Y / state Z (etc)', where X, Y, Z stand for phenomenal primitives. Although a phenomenalist might be able to inductively infer causal and probabilistic relations between different phenomenal states and between phenomenal states and states of the world, its introspective system itself reveals nothing but the instantiation of phenomenal primitives.

A *neuralist*, on the other hand, possesses an introspective repertoire derived from neuroscience. When a neuralist introspects, it characterizes its current internal states in the way Paul Churchland imagines future humans might do, in terms of such things as '[d]opamine levels in the limbic system, the spiking frequencies in specific neural pathways, resonance in the *n*th layer of the occipital cortex, inhibitory feedback to the lateral geniculate nucleus, and countless other neurophysical niceties' (Churchland, 1985, p. 16). Churchland thinks that our introspection is flexible enough to adopt this repertoire, but, whether or not he is right about this, it is certainly a possible repertoire, different from the one we currently use.

Finally, a *controller* is a creature whose introspective repertoire characterizes its internal states merely as *capacities to control* features of the world. While a human might introspect the belief that 3336 is the pin-code of their credit card, a controller will simply introspect its capacity to pay and withdraw money. While a human might realize that they are extremely afraid of the aggressive stranger threatening them, a controller would notice a loss of the capacity to walk straight or talk back, while the capacity to flee remains. That is, a controller introspects nothing but forms of relationally conceived *know-how*.

Whether such repertoires could be employed by *efficient* forms of introspection is an open question, but these examples serve to highlight the possibility of forms of introspection radically different from our own.

	Human (plausibly)		Chimpanzee (plausibly)		Phenomenalist		Neuralist		Controller	
	Grp	Cat	Grp	Cat	Grp	Cat	Grp	Cat	Grp	Cat
Direction of fit (MtW vs WtM)	✓	✓	✓	✓	?	✗	?	✗	?	✗
Perceptual vs cognitive	✓	✓	?	?	?	✗	?	✗	?	✗
Relationality	✓	✓	✓	✓	?	✗	?	✗	?	✓
Intentionality	✓	✓	✗	✗	?	✗	?	✗	?	✗
Phenomenality	?	✓	?	?	?	✓	?	✗	?	✗

Table 1: Proposed classification of introspective repertoires

Grp = Grouping, Cat = Categorization. We place a question mark in the phenomenality grouping for human introspection, since, as illusionists, we doubt there is a real grouping corresponding to the categorization. The chimpanzee introspective repertoire is based on elements discussed later in Section 4. We place question marks in the introspective grouping columns of our three imaginary creatures, since we defined these creatures in terms of how they *characterize* their introspective groupings, leaving it open whether their characterizations are accurate and the groupings useful.

2.3 Unity as a dimension of possible introspection

Other variations to consider when exploring the space of possible forms of introspection are the *range* of introspective devices possessed by a cognitive system and the degree of *unity* these devices possess.

Think again about how we could give our minimal mind introspective capacities. We could begin by installing a single introspective device, employing a single type of process and a limited number of representations of a similar type. Its introspection would thus be strongly *unified* in devices and repertoire. We could then add some diversity by equipping the single device with a wider repertoire, containing representations of different types. We could further increase the diversity by adding additional introspective devices, which rely on different types of process and have increasingly diversified repertoires.

Now consider what happens to the outputs from this collection of introspective devices. We could add a local workspace dedicated to unifying and coordinating the outputs and creating compound representations for input to the wider cognitive system. Or we might let the devices run in parallel and directly influence the rest of the system without systematic unification or coordination. If the different mechanisms represent different kinds of mental states, they might do this without impinging on each other's territory (thus achieving a sort of 'negative' coordination), but if their targets overlap, they might compete for cognitive and behavioural influence.

We thus have a scale of variation, running from extreme unity, where there is only a single device with a restricted repertoire, to extreme disunity, where there are many uncoordinated devices with diverse repertoires.

This dimension is reflected in various competing views of human introspection. While most theorists have assumed that human introspection relies on just one type of mechanism, or on a small number of coordinated ones, others have argued that it is widely fractionated (Hill, 2011; Prinz, 2004) or radically disunified (Schwitzgebel, 2012). From our perspective we can see all these approaches as describing forms an introspective system might take. The unity/disunity axis adds one more dimension to the space of possible introspections, though at a higher level than that of individual devices.

2.4 From a map to a programme

We have started to map the space of possible introspection, listing three dimensions along which introspective devices may differ, five features by which to characterize their repertoires, and a measure of the unity of a system's introspective capacities. Thus, using D_1, D_2, D_3, ... to denote introspective devices, R_1, R_2, R_3 ... to denote introspective repertoires, and u to denote a measure of unity between devices, we can define a possible form of introspection PI_x as $u\{D_iR_j, D_{i'}R_{j'} ...\}$. This is only one way to map this territory, of course, and it may not be the best. Most of this territory is *terra incognita*, and our mapping is just a first sketch.

The next step in the exploration of possible forms of introspection is to identify *interesting* regions on this map. This will be at the core of the possible introspection research programme. On a first approximation, a region is interesting if includes forms of introspection that efficiently perform the function of introspection — providing metacognitive information that can be used for online behavioural control. In Section 4 below, we shall present the programme in more detail and sketch some guidelines for conducting it. First, however, we shall discuss some objections to the view that the space of possible introspections is as large as we have suggested.

3. Are there *a priori* limitations on the space of possible introspections?

3.1. The argument from phenomenality

So far, we have been liberal in our approach, allowing that possible forms of introspection might be distributed all over the map. Some might object to this stance and argue that we can rule out some possibilities a priori. They might, for example, propose the following argument from *phenomenality*:

> **Premise 1:** For a system to introspect, it must have genuine mental states to introspect.
>
> **Premise 2:** For a system to have genuine mental states, it must have phenomenally conscious states.[5]
>
> **Premise 3**: Phenomenally conscious states are introspectively presented *as they really are* to the creature who has them.[6]
>
> **Conclusion**: Any system that introspects will be introspectively presented with its phenomenally conscious states *as they really are*.

This conclusion restricts the space of possible introspections, ruling out forms of introspection which do not employ phenomenal concepts or fail to apply them correctly.

We are not impressed by this argument. Premise 2 is not easy to defend; many philosophers would deny it — most notably, illusionists about phenomenal consciousness, such as ourselves (Frankish, 2016; Kammerer, 2021). Things are even worse for Premise 3,

[5] As John Searle puts it, ‘mental phenomena are essentially connected with consciousness’ (Searle, 1992, p. 20).

[6] This is sometimes called the *Revelation Thesis*. Both proponents and detractors have claimed that it is central to our conception of phenomenal states (Goff, 2017; Lewis, 1995).

which many naturalistic philosophers would deny.[7] Finally, one could accept all the premises but interpret the conclusion as a mere semantic point. Even if 'genuine' mentality requires phenomenality and the introspective presentation of it, we could simply drop the claim that introspection must target mental states *in this sense of 'mental'*. We could introduce a new term, 'mental*', where mental* states are ones that are functionally similar to genuine mental states but not phenomenal. We could then refocus our inquiry on *introspection**, defined in the same way as introspection, except that 'mental' is replaced with 'mental or mental*'. Thus, we invite anyone convinced by the argument from phenomenality to replace all occurrences of the word 'introspection' in this paper with 'introspection*'. The interest and value of the inquiry would not be deeply affected.

Finally, even if we focus on 'genuine' mentality, a problem remains. For the sense of 'introspection' in which Premise 3 is plausible is not the one assumed in our discussion of possible forms of introspection. It must correspond to a form of primitive, immediate introspective awareness — some form of 'intrinsic subjectivity' (Frankish, 2016, 2019) — which is not mediated by introspective representations and does not support the capacities for control that such representations afford. (Acquaintance, which we are treating as a limit case of introspection, is often thought of in this way.) Otherwise, creatures without introspective capacities could not be phenomenally conscious, which is not something most people are ready to accept. Thus, the proponent of the argument cannot rule out the possibility of a 'qualiablind' creature, which has phenomenal states and immediately introspects them but

[7] For an overview of reasons, both theoretical and empirical, for doubting that introspection always presents conscious experiences accurately, see Schwitzgebel, 2014, Section 4.

does not form introspective representations or display any of the cognitive capacities such representations would afford, including the ability to make judgements about its own experiences. This suggests that, from a cognitive perspective at least, 'genuine' mentality is not theoretically interesting.

Such considerations may not completely undermine the possibility of immediate introspection, but they do license empirically oriented researchers to ignore it and focus instead on the mediated, representation-involving kind, and the phenomenality argument does not restrict the space of possible introspective systems of that kind.

3.2. The argument from transcendental subjectivity

Another possible a priori argument limiting the space of possible introspections is from what we shall call *transcendental subjectivity*. It runs as follows:

> **Premise 1:** For a system to introspect, it must have genuine mental states to introspect.
>
> **Premise 2:** For a system to have genuine mental states, it must be able to judge of its mental states 'These are my thoughts'.[8]
>
> **Premise 3:** To be able to judge 'These are my thoughts', a system must have the concept of self and the concept of thought.

[8] This is inspired by Kant's remarks on the necessity of the 'I think', which a subject must be able to join to any of their representations in order for them to be genuinely *their* representations — *their* mental states (Kant, 1998, pp. B131-132).

> **Conclusion:** Any system that introspects will possess an introspective repertoire including the concept of self and the concept of thought.

This conclusion also restricts the range of systems that qualify as genuinely introspective, ruling out all non-conceptual representational devices as well as conceptual devices with certain repertoires, even if these devices serve to represent states functionally similar to mental ones in a manner allowing for online behavioural control.

This argument might capture an intuitive line of thought, but we do not think it should be used to restrict the scope of our inquiry. Again, one could accept its premises but treat the conclusion as a mere semantic point. Even if 'genuine' mentality requires mastery of the concepts of self and thought, we could drop the claim that introspection must target mental states in this sense of 'mental'. Following the same strategy as before, we could introduce a new term, 'mental**', referring to states functionally similar to genuine mental states but not requiring specific conceptual abilities for their possession. And we could refocus our inquiry on *introspection***, defined as targeting states that are either mental or mental**. Again, we are confident that the inquiry would lose little of its interest and value.

We suspect that analogous points could be made in response to other arguments for an a priori restriction on the space of possible introspections. At best they would establish merely semantic points, which we could acknowledge without changing our approach. Which forms of introspection are really possible — and interesting — thus remains an open question.

Our response to these arguments also addresses a more general objection to our approach. As noted in Section 1, some might resist our liberal definition of introspection, which did not mention various features traditionally associated with it, such as the first-personal perspective, phenomenal consciousness, the privacy of the mental, and the authority of the subject. It is true that our definition is not traditional. However, this is because it is not intended as an *analysis* either of our ordinary (or 'folk') concept of introspection or of the concept employed by philosophers of mind. It is a working definition allowing us to approach the issue of self-representing systems from an engineering perspective, unconstrained by preconceptions about how our minds grasp themselves. Those who find the traditional perspective valuable (and they may have good reasons to do so) can still contribute to our project — and find it valuable — by understanding it as concerning a theoretically interesting form of mental self-representation, such as introspection* and/or introspection**, rather than introspection in the traditional sense.

4. The possible introspections research programme

4.1. Preliminary remarks

This section will introduce the possible introspections research programme by outlining a set of provisional guidelines for exploring the multi-dimensional space of possible forms of introspection and identifying its most interesting regions. We begin with some preliminary remarks.

We said earlier that a region is interesting if it includes forms of introspection that efficiently perform the function of introspection — allowing for efficient online behavioural control. But, of course, interestingness is not an absolute matter. The efficiency, and hence interestingness, of a form of introspection will be *system-relative*. Introspective information that would promote efficient online behavioural control in one type of creature might be useless to another. Indeed, *all* forms of introspection will be inefficient for a vast range of possible minds, simply because they don't need introspection and couldn't make use of metacognitive information if they had it. We can partly address this problem by relativizing interestingness to minds that are sophisticated enough to find introspection useful, but it will still be hard to generalize about which types of introspection might prove efficient. Maybe some forms would be efficient for a wide variety of sophisticated minds, while others would be efficient for virtually none.

Moreover, regions of the space of possible introspection might be more interesting the further removed they are from the region occupied by human introspection. Discovering that there are efficient ways to introspect that are very different from our own could be particularly illuminating. We should thus expect fruitful interaction between research on actual introspection in humans and determination of interesting regions in the space of possible introspection.

Finally, note that the two stages of the project — mapping the space of possible introspection and identifying interesting regions of it — will not always be clear-cut. Mapping the space might alert us to new ways in which introspection could be useful, enriching our concept of interestingness, and evaluating the interest of a region might reveal new dimensions along which introspection could vary, leading us to revise our map. Again, the map proposed in Section 2 is only a first sketch, and we expect it to be enriched and revised.

With these remarks in place, we turn now to the exploration itself. Ways of exploring the space of possible introspections can be roughly divided into two categories: case-driven and theory-driven.

4.2. Case-driven exploration

Case-driven exploration involves surveying cases of existing introspective systems, natural and artificial, and looking for diversity among them. We can divide the exploration into three categories, focusing on introspection in humans, non-human animals, and AIs.

4.2.1 Humans

An obvious way to study the space of possible introspection is to look at how introspection varies in humans. How does human introspection vary with culture and language? How is it influenced by scientific, religious, moral, or philosophical beliefs? How is it affected by mental training techniques, such as meditation, hypnosis, and introspective training (Morris, 2021)? How is it modified by therapy, religious confession, and the consumption of psychoactive drugs? How does introspection differ between adults and children? How does it differ in people with non-typical neurological conditions, such as autism and schizophrenia?

Answering these questions will not be easy. There is little agreement on how introspection works even in the most studied ‘standard’ cases (typically, educated adult Westerners). How can variations of introspection across multiple dimensions be studied if we do not even agree on what introspection is like in a restricted range of baseline cases? One response would be to focus on coarse-grained features on which there is some agreement. For example, if adult Westerners generally agree that introspection characterizes some types of mental state as intentional, then we could ask if this feature is always present. Is it culturally specific? Is it affected by meditation? And so on. Another option would be to make some strong but provisional assumptions about the baseline cases for heuristic purposes.

A second difficulty arises from the fact that, although introspection itself has been relatively neglected by researchers, closely related phenomena have been extensively studied by psychologists, anthropologists, linguists, and others. In particular, there has been much work on the psychological concepts and principles assumed in everyday social interaction — ‘theory of mind’ or ‘ToM’ — and how they vary across cultures. The problem comes from the fact that, while the ToM and introspection are plausibly related, so that studying one should provide some information about the other, there is no agreement as to exactly *how* they are related. Is ToM dependent on introspection? Is introspection just self-applied ToM? This means that study of the variations of human introspection may involve reinterpreting data collected for other purposes and from different perspectives.

A third difficulty stems from the complex interrelation between first-order mental states, introspective processes, and behavioural responses. Suppose we ask people to report on their current mental states while under controlled conditions, varying the conditions along lines such as those previously mentioned (culture, language, meditation, etc.). Variation in reports could be due to some or all of the following: (a) variation in pre-introspective factors, such as the first-order mental states being introspected, (b) variation in the introspective processes themselves, and (c) variation in post-introspective factors, such as background beliefs and linguistic competence, which influence how the outputs of introspection are reported. (For some introspective devices, linguistic abilities and background beliefs might be involved in the introspective process itself, making the relation even more complex.) This means that theories of introspection are underdetermined by data about responses, and researchers will need to devise experimental protocols that maximize the chances of detecting genuinely introspective variation.

We shall now say something about introspective variation in humans, focusing on two potential sources of variation: culture (including language) and meditative practice.

People’s beliefs and reports about their own minds vary with cultural and linguistic factors. Scientific and philosophical theories of the mind have been highly diverse throughout history and between cultures. ToM is possibly more stable and may have a universal core centred on belief-desire psychology, but it may still be cross-culturally variable outside the core, for example in the classification of emotions or the thought/feeling distinction (Lillard, 1998). Some linguists claim there are partial variations in mentalizing language, with some terms, such as ‘want’, ‘think’, ‘know’, and ‘feel’, being cross-linguistically translatable, and others, such as ‘experience’, being more parochial (Wierzbicka, 2019). However, it is not clear that these variations in mentalistic beliefs and language correspond to variations in introspection and introspective repertoires. Introspection might be a stable, universal process, which interacts with cultural and other factors to generate ToM, and variations in ToM might reflect differences in those factors.[9] More generally, studying what people are inclined to *say* about minds (which is what the study of ‘intuitions’ about the mind comes down to)[10], or how they attribute mental states to others (which is what much of the study of ToM and ‘mindreading’ is about)[11] is unlikely to provide data that warrants firm conclusions about the extent of introspective variation. Here the second and third problems described above present themselves again.

How could researchers address these problems? A starting point would be to focus on tasks designed to be as introspective as possible. Methods for the careful collection of introspective data have been developed, most notably by Russell Hurlburt (the Descriptive Experience Sampling method; e.g., Hurlburt & Schwitzgebel, 2007; Hurlburt, 2011) and by Pierre Vermersch (the phenomenology-inspired ‘explicitation interview’ method; Vermersch, 1994). However, these methods have focused more on the introspected mental states than on the introspective process itself. One way to move forward would be to use these methods to collect introspective data from subjects carrying out first-order tasks — say, perceptual ones.

[9] See the CIAO (culture, introspection, analogy, ontogeny) approach to ToM developed by Angeline Lillard (Lillard, 1999).

[10] See, e.g., Knobe & Prinz, 2008.

[11] For overviews, see Marraffa, 2021 and Whiten, 2006.

If we find differences in subjects' introspective reports on their first-order mental states which do not correspond to differences in their performance on the first-order tasks, then this would be evidence that the subjects differ specifically in their introspective processes.

Stable individual introspective differences across perceptual tasks have already been observed for some simple introspective tasks, such as confidence estimates (e.g., Song et al., 2011), but we shall need to focus on more complex introspective tasks if we are to detect potential variation in the *content* of introspective representations. In co-authored work, one of us has used questionnaires and interviews to study introspective differences in complex judgments (regarding the overflowing character of visual experience) without differences in perceptual performance (Cova et al., 2020). This method could be employed in cross-cultural and cross-linguistic studies to uncover potential variation in introspective representations.

Introspective reports may also be affected by meditative practices. There are many meditative techniques, involving different capacities (sensory, cognitive, etc.), different focuses (body, breath, feelings, mental images, etc.), and different theoretical traditions (Buddhism, Vedanta, Yoga, Taoism, Sufism, etc.). Some have claimed that all these techniques converge, leading to a condition where the meditator apprehends their inner state as one of 'pure consciousness' devoid of any specific content, variously called 'emptiness', 'suchness', or 'being' (Shear, 2007, p. 700). This state of pure consciousness is taken by some meditators to be one that underlies ordinary experience, even though it usually goes unnoticed (Shear, 2007, p. 701). If this is correct, then intense meditation can modify human introspection, supplementing its repertoire with a new mental representation which does not characterize its target as having any specific phenomenal or intentional character. This new introspective representation could then, perhaps, be applied to other aspects of one's mental life, and we could investigate this by studying how meditative practice affects introspective reports in non-meditative contexts. Such an intra-personal approach should help to minimize the role of non-introspective factors in introspective variation.

Of course, it may not be the case that different meditative practices converge. Other researchers have claimed that different traditions lead their practitioners to different introspective outputs and taxonomies, perhaps because of the different theoretical assumptions adopted (Garfield, 2015, pp. 184–186). If this is so, then mapping these different taxonomies should itself teach us something about possible variation in introspective repertoires and the extent to which human introspection is flexible and conceptualized.

Finally, research into the effects of meditation on introspection need not be limited to studying the reports of expert meditators in established traditions. Even if traditional forms of meditation do not modify introspection, it may be that other forms could be devised which do. One could then try to experimentally modify human introspection by using such techniques and examine what variation results.

4.2.2 Animals

Another case-driven way to explore introspective variation is by studying non-human animals (henceforth ‘animals’). Do some animals introspect, and if so, how do their introspective systems differ from ours? Where do they lie in the space of possible introspective systems? It makes sense to begin with the animals likeliest to possess introspective abilities, such as primates (notably apes), and other mammals, such as dogs, elephants, and dolphins. However, the most interesting variants might be found on more distant branches of the evolutionary tree, where introspection, if it occurs, would be less likely to stem from commonly inherited capacities. The discovery of forms of introspection in intelligent birds, such as corvids, or, even better, in intelligent invertebrates, such as cephalopods (cuttlefish, squid, and octopuses), would be particularly exciting and informative.

The study of introspection in animals will face similar difficulties to those facing its study in humans, and two factors specific to animals exacerbate the underdetermination problem. First, animals cannot produce linguistic reports or respond to explicit task instructions, making it harder to interpret their behavioural responses. Second, in most cases, the hypothesis that animal responses are not sensitive to introspective factors will always be a serious contender.

We shall now look at some of existing research on animal cognition that might prove relevant, beginning with work on theory of mind in animals**.** Since the late 1970s, extensive research has been conducted on whether animals, in particular chimpanzees, possess some understanding of mental states and their influence on behaviour — a theory of mind (ToM) (Premack & Woodruff, 1978). Evidence indicated that chimpanzees have some ability to represent others' mental states (Call & Tomasello, 2008), and there was some, weaker evidence for the possession of a ToM by other primates (Flombaum & Santos, 2005), some species of birds (Dally, 2006), and even cephalopods (Mather & Dickel, 2017). But though chimpanzees can represent *factive* mental states, involving knowledge of, perception of, or desire for an actual state of affairs or present object, the evidence indicates that they cannot represent non-factive ones. They do not seem able to represent that others have false beliefs (that is, beliefs about non-actual states of affairs) or desires for things that are not present, and thus appear to lack the concepts of belief and desire that we have.[12]

Even if chimpanzees do possess a simple ToM, it does not follow that they can apply it to themselves (Focquaert et al., 2008). If they can, however, they would employ an introspective repertoire different from the typical human one, involving representations of mental states as *relational* (constituted by relations to objects or actual situations) but not *intentional* (bearing intentional content). This remains a possible form of introspection even if chimpanzees do not in fact employ it.

Research on animal ToM offers only tentative support for claims about animal introspection. Things look different when we turn to research on animal *metacognition*. For decades now (see Beran, 2019 for an overview), researchers have studied the ability of animals to represent aspects of their own cognitive processes — an ability usually referred to as 'metacognition' and close to what we call 'introspection'. Using techniques such as betting paradigms (typically designed as to make it advantageous *not* to bet in situations of uncertainty), researchers have investigated the capacity of various animals (apes, rhesus monkeys, dolphins, as well as rats, cats, and pigeons) to represent features of their own mental states, including their degrees of confidence, their possession of items of knowledge, and their capacity to memorize facts. It is still unclear whether animals really possess such capacities, and some researchers have argued that their behaviour can be interpreted in purely first-order terms, without reference to metacognitive processes (Carruthers, 2008; Carruthers &

[12] More recent work suggests that in certain settings chimpanzees *can* represent that others have false beliefs (Krupenye et al., 2016).

Williams, 2019). So far, however, the debate has centred on whether animals possess metacognition at all; less attention has been given to characterizing the cognitive architectures that might support their putative metacognitive abilities. Such research could provide important insights into possible introspective devices and repertoires.

4.2.3 AI systems

AI should provide us with an increasing number of case studies in forms of introspection. Sophisticated artificial agents will need to monitor their own internal states for the purposes of self-regulation, and they will increasingly need to share information about their internal states with other agents (for an early speculative exploration, see McCarthy, 2000, and for a more recent one, see Fleming, 2021, ch.10). AIs designed to interact with humans, such as care robots, will need to employ mental concepts close to our own. However, those that communicate primarily with other artificial agents will not be so constrained, and if they evolve self-monitoring and self-reporting mechanisms through a process of autonomous self-improvement, they may develop forms of introspection very different from ours. This could allow us to explore interesting unknown regions of the space of possible introspections. (If, on the other hand, such artificial agents systematically evolve introspective devices and repertoires similar to ours, this would also teach us much about the structure of the space and the rarity of interesting regions within it.)

Forms of introspection (or metacognition) in artificial systems already exist. Some aim to simulate or replicate features of human introspection, but these may still have interest for our programme, since there may be interestingly distinct ways of replicating human representational processes. For instance, Sloman has pointed to significantly different ways in which an artificial system could represent states that themselves have representational content (Sloman, 2011, pp. 312–313).

More insights could be gained from the study of AI systems with metacognitive capacities that were not designed to mimic human introspection. Many cognitive architectures are already capable of monitoring themselves; they can, for example, represent the availability of internal resources or the presence or absence of knowledge regarding a given task (Kotseruba & Tsotsos, 2020, pp. 56–57). And we can expect to see the development of more complex introspective devices for AIs which need to communicate about their internal states and must therefore *repackage* information about them in an easily transmissible and interpretable format.

Some interesting features of artificial metacognition are already evident. Researchers have noted that artificial systems that are vulnerable to external attack and intrusion may benefit from having multiple independent self-monitoring mechanisms directed at the same processes (Kennedy, 2011; Sloman, 2011, pp. 311). Massively fractionated and distributed forms of introspection might thus be efficient for artificial systems, if not for biological ones, which are not vulnerable to attack in the same way. This suggests that interesting possible forms of introspection may lie at locations on the unity axis far from that occupied by human introspection.

4.3. Theory-driven exploration

Theory-driven exploration investigates the space of possible introspection by reflecting on theoretical considerations about the nature and function of introspection. The a priori argument examined in Section 3 attempted to use theoretical considerations to limit the space of possible introspection. However, theoretical approaches can also be pursued in a more positive manner. We can divide theoretical approaches into *system-based* and *topic-based*.

System-based exploration starts with models of actual introspective systems, and then extrapolates from them, asking what variations of the model are permitted theoretically and what consequences these variations would have for the functioning of the system. Given a model of an introspective system composed of a number of interconnected devices each with certain features and a certain introspective repertoire, we might imagine: (a) adding or removing introspective devices, (b) modifying a given device, perhaps along one of the dimensions described earlier, (c) changing an introspective repertoire, either by adding or removing representations or by altering its structure, and (d) modifying the way in which the various devices are coordinated.

Theory-driven exploration of this kind is continuous with case-driven exploration, since it begins with a model of some actual introspective system, but it may take us well beyond the original case, and simple models may lead us to explore complex theoretical possibilities. Suppose we find that octopuses, with their massively distributed nervous systems, have self-monitoring mechanisms that are very different from ours. These might in themselves be simple and relatively uninteresting from a theoretical point of view. But taking our model of them as a starting point, we might enrich and complicate it, devising models of interesting possible forms of introspection which, so far as we know, are not exemplified in reality. In this way, case-driven studies could provide us with new theoretical building blocks, enabling us to construct theoretically interesting models.

Moreover, once we have identified an interesting possible form of introspection in this way, we could try to simulate it in software or implement it in a robotic system. We might even implement such systems, or elements of them, in the form of prostheses that could be coupled with our own cognitive systems. In this way, the exploration of possibilities might lead to the development of actual technology which enriches our native introspective capacities.

A system-based theoretical approach will become easier as we make progress in modelling actual cases of introspection. It will also require a form of imagination close to that required in engineering, and it might also benefit from wilder, science-fictional speculations.

The other kind of theory-driven exploration is *state-based*. This starts, not with a model of an introspective *system*, but with a model of the introspected *mental states.* Consider, for example, the actual mental states targeted by human introspection — the states we call ‘beliefs’, ‘desires’, ‘perceptions’, ‘emotions’, and so on. Cognitive scientists aim to model these states, and they have developed a wide range of models belonging to a variety of theoretical frameworks (computational, connectionist, dynamical, enactive, predictive processing, etc.). Now, in pursuing this modelling they might discover new patterns in the structure and dynamics of mental states, including ones that would be easily detectable, easily modifiable, and highly predictive. Such patterns are precisely the ones that interesting possible forms of introspection would target, since by tracking them a cognitive system could enhance its capacity for efficient online behavioural control. And the more these patterns diverge from those tracked by our own form of introspection, the more the forms of introspection that track them would diverge from ours.

Note that the patterns that are interesting from this perspective would not necessarily be the same as those that are interesting from the perspective of cognitive science more broadly. When engaged in general theorizing, cognitive scientists look for patterns that help them explain a system’s activity and responses. Focusing on introspection, however, they would look for patterns it would be useful for the *system itself* to represent in order to enhance its capacity for self-control.

It may be that we shall not discover such patterns, and that the only easily detectable, easily modifiable, and highly predictive patterns in our mental states are those picked out by our existing introspective repertoires — the ones that evolutionary and cultural processes have already homed in on. But we should not assume that this is so. Evolutionary and cultural processes do not always find optimal solutions, and it would be strange if there were no other relevant unexploited information nested within the structure and dynamics of our mental states. Again, this line of inquiry might result in the development of technology that enhances our own introspective capacities.

Conclusion

What forms could introspection take? The question is under-explored, and we believe that dedicating some time to it could provide important insights for the study of minds, both human and non-human. We have made the case for the interest of the question and sketched a research programme for its further exploration. We hope others will take up the challenge.

References

Anscombe, G. (1957). *Intention*. Blackwell.

Armstrong, D. (1980). *The nature of mind and other essays*. Cornell University Press.

Bartlett, S. & Wong, M. L. (2020). Defining lyfe in the universe: From three privileged functions to four pillars. *Life*, *10*(4), 42.

Bayne, T. & Montague, M. (Eds.). (2011). *Cognitive phenomenology*. Oxford University Press.

Beck, J. (2018). Marking the perception–cognition boundary: The criterion of stimulus-dependence. *Australasian Journal of Philosophy*, *96*(2), 319–334.

Beran, M. (2019). Animal metacognition: A decade of progress, problems, and the development of new prospects. *Animal Behavior and Cognition*, *6*(4), 223–229.

Brentano, F. (1874/2015). Psychology from an empirical standpoint. Routledge.

Byrne, A. (2018). *Transparency and self-knowledge*. Oxford University Press.

Call, J. & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *TRENDS in Cognitive Sciences*, *12*(5), 187–192.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language, 23*(1), 58–89.

Carruthers, P. (2011). *The opacity of mind*. Oxford University Press.

Carruthers, P. & Williams, D. M. (2019). Comparative metacognition. *Animal Behavior and Cognition*, *6*(4), 278–288.

Chalmers, D. (2003). The content and epistemology of phenomenal belief. In Q. Smith & A. Jokic (Eds.), *Consciousness: New philosophical perspectives* (pp. 220–272). Oxford University Press.

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *The Journal of Philosophy*.

Cova, F., Gaillard, M., & Kammerer, F. (2020). Is the phenomenological overflow argument really supported by subjective reports? *Mind and Language*.

Dally, J. M. (2006). Food-caching western scrub-jays keep track of who was watching when. *Science*, *312*(5780), 1662–1665.

Dennett, D. (2017). From bacteria to Bach and back. Norton.

Dretske, F. (1981). Knowledge and the flow of information. MIT Press.

Dunne, J. D., Thompson, E., & Schooler, J. (2019). Mindful meta-awareness: sustained and non-propositional. *Current Opinion in Psychology, 28,* 307–311.

Fleming, S. M. (2021). *Know thyself: the science of self-awareness* (First edition). Basic Books.

Flombaum, J. I. & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology*, *15*(5), 447–452.

Focquaert, F., Braeckman, J., & Platek, S. M. (2008). An evolutionary cognitive neuroscience perspective on human self-awareness and theory of mind. *Philosophical Psychology, 21*(1), 47–68.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71.

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

Frankish, K. (2019). The meta-problem is *the* problem of consciousness. *Journal of Consciousness Studies*, *26*(9–10), 83–94.

Frith, U. & Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind and Language*, *13*(1), 1–22.

Garfield, J. L. (2015). *Engaging Buddhism: Why it matters to philosophy*. Oxford University Press.

Gertler, B. (2012). Renewed acquaintance. In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness* (pp. 89–123). Oxford University Press.

Goff, P. (2017). *Consciousness and fundamental reality*. Oxford University Press.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*, 1–14.

Graziano, M. (2013). *Consciousness and the social brain*. Oxford University Press.

Hill, C. (2011). How to study introspection. *Journal of Consciousness Studies*, *18*(1), 21–43.

Hurlburt, R. T. (2011). Investigating pristine inner experience: Moments of truth. Cambridge University Press.

Hurlburt, R. T. & Schwitzgebel, E. (2007). *Describing inner experience? Proponent meets skeptic*. MIT Press.

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

Kant, I. (1998). *Critique of Pure Reason*. (P. Guyer & A. W. Wood, Eds.) Cambridge University Press.

Kennedy, C. M. (2011). Distributed metamanagement for self-protection and self-explanation. In M. T. Cox & A. Raja (Eds.), *Metareasoning* (pp. 233–248). MIT Press.

Knobe, J. & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the cognitive sciences*, *7*(1), 67–83.

Kotseruba, I. & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17–94.

Kriegel, U. (2019). Phenomenal intentionality and the perception/cognition divide. In A. Sullivan (Ed.), *Sensations, thoughts, language: Essays in honor of Brian Loar* (pp. 167–183). Routledge.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308).

Levin, J. (2007). What is a phenomenal concept? In T. Alter & S. Walter (Eds.), *Phenomenal concepts and phenomenal knowledge: New essays on consciousness and physicalism* (pp. 87-110). Oxford University Press.

Lewis, D. (1995). Should a materialist believe in qualia? *Australasian Journal of Philosophy, 73*(1), 140–44.

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin, 123*(1), 3–32.

Lillard, A. (1999). Developing a cultural theory of mind: The CIAO approach. *Current Directions in Psychological Science, 8*(2), 57–61.

Marraffa, M. (2021). Theory of mind. In *Internet encyclopedia of philosophy*. https://iep.utm.edu/theomind/

Mather, J. A. & Dickel, L. (2017). Cephalopod complex cognition. *Current Opinion in Behavioral Sciences*, *16*, 131–137.

McCarthy, J. (2000). Making robots conscious of their mental states. In K. Furukawa, S. Muggleton & D. Michie (Eds.), *Machine intelligence 15: Intelligent Agents* (pp. 3–17). Oxford University Press.

Morris, A. (2021). Invisible gorillas in the mind: Internal inattentional blindness and the prospect of introspection training [Preprint]. PsyArXiv, 26 Sept. 2021. Web.

Nichols, S. & Stich, S. (2003). How to read your own mind: A cognitive theory of self-consciousness. In Q. Smith & A. Jokic (Eds.), *Consciousness: New philosophical perspectives*. Oxford University Press.

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526.

Prinz, J. (2004). The fractionation of introspection. *Journal of Consciousness Studies, 11*(7–8), 40–57.

Schwitzgebel, E. (2012). Introspection, what? In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness*, (pp. 29–48). Oxford University Press.

Schwitzgebel, E. (2014). Introspection. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Summer 2014 Edition). http://plato.stanford.edu/archives/sum2014/entries/introspection/

Searle, J. (1992). *The rediscovery of the mind*. MIT Press.

Shear, J. (2007). Eastern methods for investigating mind and consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* (pp. 697–710). Blackwell Publishing.

Sloman, A. (2011). Varieties of metacognition in natural and artificial systems. In M. T. Cox & A. Raja (Eds.), *Metareasoning* (pp. 307–322). MIT Press.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787–1792.

Spaulding, S. (2015). On direct social perception. *Consciousness and Cognition*, *36*, 472–482.

Vermersch, P. (1994). *L'entretien d'explicitation*. ESF.

Whiten, A. (2006). Theory of mind. In L. Nadel (Ed.), *Encyclopedia of cognitive science.* John Wiley.

Wierzbicka, A. (2019). From 'consciousness' to 'I think, I feel, I know'. *Journal of Consciousness Studies, 26*(9-10), 257-269.