Heather Browning[1]
and Walter Veit[2]

Studying Introspection in Animals and AIs

Abstract: *The study of introspection has, up until now, been predominantly human-centric, with regrettably little attention devoted to the question of whether introspection might exist in non-humans, such as animals and artificial intelligence (AI), and what distinct forms it might take. In their target article, Kammerer and Frankish (this issue) aim to address this oversight by offering a non-anthropocentric framework for understanding introspection that could be used to address these questions. However, their discussions on introspection in animals and AIs were quite brief. In this commentary, we will build on their suggestions to offer some methodological guidance for how future research into introspection in animals and AIs might proceed.*

Keywords: introspection; animal minds; artificial intelligence; evolution; design.

1. Introduction

The ability to introspect has received much attention among both philosophers and scientists. However, as with other mental phenomena such as consciousness (see Veit, 2023), research on introspection has largely only investigated its instantiation in humans, with very

Correspondence:
Email: drheatherbrowning@gmail.com

1 University of Southampton, UK.

2 University of Bristol, UK, & Munich Center for Mathematical Philosophy, LMU, Germany.

DOI: 10.53765/20512201.30.9.063

little attention given to the distinct forms of introspection we might find in non-human systems, whether animals or AIs. This narrow focus is unfortunate, since it means that we could at best only derive a theory of *human* introspection, rather than a more general theory describing a broader natural phenomenon that may be shared by both living and artificial systems.

In their target article, Kammerer and Frankish (this issue; henceforth K&F) lament this neglect in the literature, aiming to offer a tentative research programme for investigating the different forms that introspection may take in varied cognitive systems — both actual and possible. As philosophers working on different kinds of minds, we are naturally enthused about their proposal. Too often, philosophers and scientists have confused the question of how mental processes — such as consciousness, introspection, memory, and the like — work in humans with the more general question of how we can understand them as broader natural phenomena. There has been a widespread failure to recognize that this is a different research question entirely, an oversight that K&F thankfully acknowledge. Additionally, understanding introspection as a natural phenomenon requires us to look not only at its actual extant instances, but also its possible ones; similar to the way in which understanding of living systems also requires exploring the other forms or body plans life may take. As one of us has argued recently, theories of mental processes have to account for the 'full diversity and complexity' of mental phenomena across species as well as a within our own (Veit, 2023, p. 41); which is echoed in the argument by K&F that we require a non-anthropocentric approach to the mind that recognizes the 'diversity and complexity of terrestrial minds' (p. 19). We could thus hardly be more in agreement with their general approach and share their prediction that their tentative research programme will also help us to better recognize and understand neurodiversity even within our own species.

Our goal in this article will thus not be to offer some general criticism of their proposal but rather to 'take up the challenge' they extend, to investigate the diversity of forms introspection is likely to take in non-human minds (K&F, p. 45). While they begin their paper by emphasizing how useful their framework will be for the study of introspection in animals and AIs, they have not been able to dedicate much space to detailing what this might look like. Here, we use this paper to expand on their discussion and offer some methodological suggestions as to how research into non-human introspection may proceed.

This article is structured as follows. In Section 2, we will discuss some possible methods, including use of evolutionary thinking, for studying introspection in animals. In Section 3, we will examine how these approaches to the study of introspection in animals can be exported to the AI case, and some of the additional methodological challenges this raises. Finally, in Section 4, we will conclude the discussion and offer some further suggestions for a truly comparative study of introspection across a variety of minds.

2. Introspection in Animals

The project proposed by K&F follows an emerging tradition within the sciences of animal cognition more generally, which seeks to explore the dimensions and variety within a cognitive trait such as consciousness (Andrews, 2022; Birch, Schnell and Clayton, 2020; Veit, 2023) or affect (Browning, 2022), rather than simply map its presence or absence. Such a research programme can tell us more about what it is like to be different types of animal, and adding an understanding of introspective processes can only help deepen such an understanding. Further, a multidimensional approach helps to better integrate empirical research where these concepts are operationalized in different ways which has previously caused many needless conceptual disputes.

However, K&F are right to emphasize the methodological challenges in researching introspection in animals. After all, as they point out, this research is difficult even in humans, where we have access to the 'gold standard' measure of verbal self-report. There is the problem of underdetermination, where observed behavioural variation could be explained not just by variation in introspective processes but also by variation in first-order mental states or production of behavioural responses. While we agree that this makes research into animal introspection more difficult, we think there are some potential ways forward, which we will outline here.

To begin with, there is the potential of using modified paradigms of self-report. Though the authors take it for granted that self-report is not possible in animals, we suggest that there are indeed some methods that may count as self-report (or at least, *quasi*-self-report) that could be used with animals. These include what we shall call 'symbolic direct reporting', 'non-symbolic direct reporting', and 'indirect reporting'. Let us briefly describe each.

While it is true that most animals are incapable of using symbolic language for self-report, there are some possible exceptions (Pepperberg, 2017; Péron, 2012). Some animals, such as great apes (Fouts and Mills, 1997; Gardner and Gardner, 1969; Miles, 1990; Patterson, 1978; Premack, 1971; Rumbaugh, 1977; Savage-Rumbaugh, 1986; Terrace *et al.*, 1979), parrots (Pepperberg, 2000), and dogs (Hunger, 2021; Rossi and Ades, 2008), have been trained to use symbols to name objects, properties, and in some cases seemingly their own internal states, such as desires (Péron, 2012). While this work has remained controversial regarding whether or not it counts as language use, this is not important for its use in the context of the type of research we are considering here. What is relevant is that it shows some animals are capable of making direct reports. Although there is insufficient work to determine yet the limits of this capacity, it is possible that an animal may be able to use a symbol to 'report' its experience of some mental state or another — whether first-order experiences or meta-representations — regardless of whether or not it can produce more complex linguistic utterances. However, we accept that this sort of 'verbal' self-report is limited only to single-word reports, rather than more complete descriptions of perceived interior states and thus would require more careful experimentation to draw out meaningful conclusions about introspective processes.

There are also more basic types of non-symbolic direct self-report that have been used with animals, without the need for learning or producing symbols. Some animals have been trained to perform a specific behaviour to indicate their affective state, which we take to be a form of introspective report. For example, pigs were able to indicate (through pressing a lever) whether or not they felt anxiety and could do so when the feeling was induced through a range of different conditions, such as pharmacological manipulation, presence of novel objects, or anticipated electric shocks (Carey, Fry and White, 1992; Carey and Fry, 1993; 1995). In cases such as these where it seems animals are able to produce a deliberate response when experiencing a specific mental state, this can provide information about their knowledge of their own mental states, such as the range of states they are capable of introspecting and reporting on.

Lastly, there are forms of indirect self-report that could be used to infer introspective processes. Examples of these can be seen in metacognition studies on animals (Beran, 2019), as were discussed by K&F, where (e.g.) decisions to continue (or not continue) a task can be seen as reflecting the animal's perception of their own knowledge,

or confidence. While the authors are correct that this work has been controversial so far, we are confident that progress could be made with careful controls and repetitions in different modalities. Thinking about the likely behavioural effects of different types of introspection (and through mapping some of these in humans), we could construct a range of tasks that allow indirect inferences about introspection. This would then require first thinking through and identifying the likely behavioural markers of possession of different introspective devices or repertoires, before devising tests to identify their presence or absence in animals. A similar approach can be seen in research into animal sentience — a similarly difficult target for scientific measurement. Here, use of a list of indicators developed from theorizing about the likely functions and effects of sentience has been used to draw inferences regarding the possession of sentience (in particular, the ability to experience pain) (Birch *et al.*, 2021; Crump *et al.*, 2022a,b; Gibbons *et al.*, 2022). Something like this could also be used as a model for investigating different forms of introspection in animals.

One concern raised by K&F is that there is more room for scepticism when interpreting animal studies. Not just because of the limitations in method, but because of different prior assumptions about cognitive capacity. While in the human case research begins from agreement that there is introspection of some type in humans and we are merely looking to map out its features and varieties, in the animal case we still need to establish whether they can introspect at all. Those with a higher degree of scepticism about complex processes in animal minds may prefer to take a 'simpler' explanation for observed behaviours. This mirrors the more general disagreements between 'romantics' and 'killjoys' about animal cognitive abilities, where background assumptions about the complexity and distribution of different abilities, and their likelihood in non-human animals, affect preferred interpretations of observed data (Andrews, 2020). However, we see that the account of introspection provided by the authors might allow for some progress here by permitting us to consider far less complex forms of introspection. In such cases, there is less reason for scepticism that animals can perform at least some of the simpler forms and thus perhaps a greater willingness to look at *what types* of introspection different animals can perform rather than becoming stuck on the question of whether they can at all.

Overcoming the twin problems of underdetermination and scepticism requires carefully designed experimental paradigms that try to control for and rule out variation resulting from anything other than

differences in introspective abilities. In particular, we think a robustness approach will be of use here. This could proceed through using multiple independent tests across different modalities, relying on varying background assumptions (see Browning, 2023, for a similar approach using robustness to validate indicators of animal welfare). There are some current good examples of this practice that could serve as models for future work, such as generalizing across perceptual or memory tests with different features, or transferring from perception tasks to memory tasks (Basile *et al.*, 2015; Brown, Templer and Hampton, 2017; Kornell, Son and Terrace, 2007; Templer and Hampton, 2012); especially when experiments are constructed with the explicit aim of ruling out alternative hypotheses. When the lines of evidence used are varied but similar results still seen, it is far more likely that they are coming from the target process (introspection) rather than alternative processes (first-order mental states or behavioural responses). In the absence of this, it might also be possible to strengthen the background theory and working assumptions to make better justified inferences about which processes are the *most likely* explanations for the observed variation, given our current best understanding of the functions and mechanisms of introspection of various types.

Finally, a research programme into animal introspection could also take a more 'bottom-up' evolutionary approach, broadening the scope to think about which ecological and life history conditions could correlate with, or even produce, different introspective features. We think it is worth emphasizing a point raised by K&F: that '*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' (this issue, p. 36). An evolutionary approach to animal minds must account for the simple fact that more complex forms of cognition will both be costly and often unnecessary. Thus, it will be very important to link introspective abilities in animals, whatever forms they may take, to the fitness benefits they would provide for these animals. Importantly, evolutionary thinking about cognitive abilities needs to be based on empirical data rather than *just* adaptationist story-telling about the possible benefits of different forms of cognition.

In response to a similar challenge for the study of animal consciousness, one of us developed the *pathological complexity thesis*: 'The function of consciousness is to enable the agent to respond to pathological complexity' (Veit, 2022a, p. 1). 'Pathological complexity' (or

alternatively, 'life history complexity') is defined in terms of the economic life history trade-offs animals are faced with during their life cycles — trade-offs that will vary substantially even for species that are morphologically very similar, if they occupy different ecological niches or pursue different life history strategies (see also Veit, 2022b,c; 2023).[3] In other words, pathological complexity is a measure of how complex the life history strategies are that different animals pursue. A deep understanding of different animals' life history strategies and the ecological challenges they face will offer us a much better framework for thinking about the evolution and fitness benefits of distinct cognitive innovations. The pathological complexity framework can be used to address the difficult question of what forms of introspection might pay off for different animals. We might, for instance, take social animals to require different capacities to solitary species, or predators to prey. Of course, this would be a bidirectional process, both informing and informed by other research into the introspective abilities of animals, as outlined above. The need to link different forms of introspection to their possible fitness benefits in the particular life history strategies of animals will help us by both (i) constraining the scope of plausible explanations for the empirical tests discussed above, and (ii) providing us with plausible hypotheses regarding the distribution of different forms of introspection across the animal branch of life, that could in turn be tested.

For instance, we hypothesize that animals with highly complex social lives should have life histories that would make greater introspective capacities worth having, in order to predict and perhaps even manipulate the behaviour of conspecifics. In corvids and apes, for instance — both taxa with complex social interactions — there has been evidence of magician-like deception (Garcia-Pelegrin *et al.*, 2022) suggesting use of introspection of the kind 'How would I be fooled myself?' in order to fool others. Further experiments could test this hypothesis and vindicate a strong relationship between introspective capacities and the social drivers of life history complexity. But perhaps more importantly, a deep understanding of the life histories of different animals will help to minimize the charge of mere speculation when it comes to investigations of the possible intro-

[3] For a discussion over why life history theory offers us the best theoretical resources for how to measure this goal-directed complexity, see Veit, Gascoigne and Salguero-Gómez (2023).

spective capacities of non-human animals by providing an answer to the question of how they would confer practical fitness benefits to the animals.

Mapping out the features of introspection across a range of animal minds could then help answer a range of additional interesting questions, such as ones about the relationships between different types of introspection and other cognitive features, such as consciousness, or types of intelligence. For instance: are some types of introspection only present in conscious animals? Does the range and complexity of introspection between species correlate with greater intelligence in other domains? The pathological complexity framework can also help answer some questions about evolution, development, and distribution of introspection. Looking for the taxonomic distribution of specific features can suggest how and when they evolved, and looking for ecological or life history correlates could help provide clues as to their functions. It will also help to inform us about the range of possible types of introspection and their mechanisms and instantiation, which can help with understanding or creating introspection in artificial systems, as we will explore in the next section.

3. Introspection in AIs

While we have described some possible ways forward for researching introspection in non-human animals, the study of introspection in AIs is potentially much more difficult. The methods we have discussed for studying animal introspection typically rely on background assumptions about evolution and life history of different species, evolved functions of introspection, or analogies with human mental processes — all of which seem unavailable (or at least, far less reliable) in the case of AIs. It will thus be harder in this case to make the same inferences from behavioural processes to features of introspection. On the other hand, our greater understanding of the mechanisms used and ability to program in features such as self-reporting may make other types of study easier.

On the surface it would seem like our proposal for the use of the pathological/life history complexity framework would be inapplicable to non-living systems since they do not have fitness by which to assess this. Yet, machines are of course even more literally products of design than are animals. Indeed, *qua* being their designers we have a much more direct understanding of the design of machines than of animals. While their goals differ, we know in fact much more about

the functional architecture of artificial systems, which provides us with an important advantage. Just as we do in the animal case, for instance, we could similarly ask what kinds of introspection it would be worth having for a machine, given its intended design.

As an example, K&F note that care robots provide an excellent case of a form of AI that would require human-like mental concepts. What is one of many proximate goals for humans (i.e. social communication), can be the ultimate goal for machines, who are obviously not 'designed' like living systems for the maximization of fitness. Elsewhere, we have explored the different ways that humans treat such social robots, arguing that humans shift regularly between a design and intentional stance[4] when dealing with them (Veit and Browning, 2023). When engaging with an entity as a social agent, future social robots will plausibly have to rely on introspective access to their own mental states to understand the behaviour of humans they are dealing with by employing the intentional stance.

Above, we mentioned that social animals such as corvids and apes have been observed to engage in deception, which could be indicative of quite complex forms of introspection in which an agent models what they themselves could be deceived by in order to deceive others. Social robots that lack similar mental states and cannot use themselves as a model to examine or predict the mental states of humans may be inherently disadvantaged compared to those who do. The intentional stance can offer practical benefits to achieving living and machine goals alike. Indeed, as one of us has argued previously in a paper with one of the authors of the target article, its origins are likely found in its application to oneself as much as others (Veit *et al.*, 2019). To design AI robots with the flexibility of human agents may thus require some form of intentional stance thinking. After all, human minds have largely been shaped by the design challenges of their social environments, so in robots designed to face similar challenges we could plausibly expect important similarities to human introspection. The worry that we cannot make any analogies with human mental processes thus seems premature.

Nevertheless, while the social robot example relates to a goal that is very similar in form to one faced by human agents, many other challenges that AIs are dealing with will have virtually nothing to do with those faced by animals (including humans). The example given

[4] For an introduction to Dennett's intentional stance, see Dennett (1988).

by K&F — that AIs would benefit from 'massively fractionated and distributed forms of introspection' (p. 43) to deal with attacks from the outside, which makes them very different from humans (or for that matter, animals) — is a particularly useful one for highlighting that AIs could have very distinct kinds of introspection. It can be a common error in cognitive science to argue from the success of an artificial system such as deep neural networks in a cognitive task, to maintaining that the human mind must operate on the same principles (Veit and Browning, forthcoming), but an error we should aim to avoid. Even human levels of performance in experimental paradigms designed for testing introspection in humans and animals might indicate something very differently in AIs. Nevertheless, since we do have a much better understanding of how computers process information than brains, this challenge could be overcome by taking a design stance that will help us to investigate what kinds of introspection would be useful for different task-specific AIs. Although they are not evolved biological systems, some of the same tools used in evolutionary thinking may still be useful for understanding the functions and constraints in artificial design.

4. Conclusion

K&F have laid the foundation for a comprehensive research programme investigating the diversity of introspective processes across human and non-human minds, including animals and AIs. In this paper, we endeavoured to contribute to their goal of a comparative study of introspection by considering some potential methods that could be used to study introspection in animals and AIs, as well as offering some tentative proposals for overcoming new challenges associated with research into non-human introspection. Necessarily, this could not here be a detailed or comprehensive analysis, but we hope to have provided a set of suggestions that can serve as the basis of future work developing more sophisticated methodologies for studying the range of introspection in non-humans. In particular, we want to emphasize that the study of introspection in animals holds a special role, allowing us to expand our methodological toolkit and practise our design thinking — essential in creating new artificial forms of introspection and examining the widely different forms of introspection that AIs may be able to achieve in the future.

References

Andrews, K. (2020) *How to Study Animal Minds*, 1st ed., Cambridge: Cambridge University Press.

Andrews, K. (2022) Does the sentience framework imply all animals are sentient?, *Animal Sentience*, **7** (32).

Basile, B.M., Schroeder, G.R., Brown, E.K., Templer, V.L. & Hampton, R.R. (2015) Evaluation of seven hypotheses for metamemory performance in rhesus monkeys, *Journal of Experimental Psychology: General*, **144**, pp. 85–102.

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

Birch, J., Schnell, A.K. & Clayton, N.S. (2020) Dimensions of animal consciousness, *Trends in Cognitive Sciences*, **24** (10), pp. 789–801.

Birch, J., Burn, C., Schnell, A., Browning, H. & Crump, A. (2021) *Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans*, London: LSE Consulting.

Brown, E.K., Templer, V.L. & Hampton, R.R. (2017) An assessment of domain-general metacognitive responding in rhesus monkeys, *Behavioural Processes*, **135**, pp. 132–144.

Browning, H. (2022) How should we study animal consciousness scientifically?, *Journal of Consciousness Studies*, **29** (3–4), pp. 12–14.

Browning, H. (2023) Validating indicators of subjective animal welfare, *Philosophy of Science*, first view, pp. 1–10. doi: 10.1017/psa.2023.10

Carey, M.P., Fry, J.P. & White, D.G. (1992) The detection of changes in psychological state using a novel pharmacological conditioning procedure, *Journal of Neuroscience Methods*, **43** (1), pp. 69–76.

Carey, M.P. & Fry, J.P. (1993) A behavioural and pharmacological evaluation of the discriminative stimulus induced by pentylenetetrazole in the pig, *Psychopharmacology*, **111** (2), pp. 244–250.

Carey, M.P. & Fry, J.P. (1995) Evaluation of animal welfare by the self-expression of an anxiety state, *Laboratory Animals*, **29** (4), pp. 370–379.

Crump, A., Browning, H., Schnell, A., Burn, C. & Birch, J. (2022a) Animal sentience research: Synthesis and proposals, *Animal Sentience*, **7** (32).

Crump, A., Browning, H., Schnell, A., Burn, C. & Birch, J. (2022b) Sentience in decapod crustaceans: A general framework and review of the evidence, *Animal Sentience*, **7** (32).

Dennett, D.C. (1988) Précis of *The Intentional Stance*, *Behavioral and Brain Sciences*, **11** (03), art. 495.

Fouts, R. & Mills, S. (1997) *Next of Kin: My Conversations with Chimpanzees*, New York: Avon Books.

Garcia-Pelegrin, E., Schnell, A.K., Wilkins, C. & Clayton, N.S. (2022) Could it be proto magic? Deceptive tactics in nonhuman animals resemble magician's misdirection, *Psychology of Consciousness: Theory, Research, and Practice*, **9**, art. 3.

Gardner, R.A. & Gardner, B.T. (1969) Teaching sign language to a chimpanzee: A standardized system of gestures provides a means of two-way communication with a chimpanzee, *Science*, **165** (3894), pp. 664–672.

Gibbons, M., Crump, A., Barrett, M., Sarlak, S., Birch, J. & Chittka, L. (2022) Can insects feel pain? A review of the neural and behavioural evidence, in Jurenka,

R. (ed.) *Advances in Insect Physiology*, vol. 63, pp. 155–229, Amsterdam: Elsevier.
Hunger, C. (2021) *How Stella Learned to Talk: The Groundbreaking Story of the World's First Talking Dog*, New York: William Morrow.
Kammerer, F. & Frankish, K. (this issue) What forms could introspective systems take? A research programme, *Journal of Consciousness Studies*, **30** (9–10).
Kornell, N., Son, L.K. & Terrace, H.S. (2007) Transfer of metacognitive skills and hint seeking in monkeys, *Psychological Science*, **18** (1), pp. 64–71.
Miles, H.L.W. (1990) The cognitive foundations for reference in a signing orangetan, in Taylor Parker, S. & Gibson, K.R. (eds.) *'Language' and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*, pp. 511–539, Cambridge: Cambridge University Press.
Patterson, F.G. (1978) The gestures of a gorilla: Language acquisition in another pongid, *Brain and Language*, **5** (1), pp. 72–97.
Pepperberg, I.M. (2000) *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*, Cambridge, MA: Harvard University Press.
Pepperberg, I.M. (2017) Animal language studies: What happened?, *Psychonomic Bulletin & Review*, **24** (1), pp. 181–185.
Péron, F. (2012) Language-trained animals: A window to the 'Black Box', *International Journal of Intelligence Science*, **02** (04), pp. 149–159.
Premack, D. (1971) Language in chimpanzee?, *Science*, **172** (3985), pp. 808–822.
Rossi, A.P. & Ades, C. (2008) A dog at the keyboard: Using arbitrary signs to communicate requests, *Animal Cognition*, **11** (2), pp. 329–338.
Rumbaugh, D.M. (1977) *Language Learning by a Chimpanzee: The Lana Project*, New York: Academic Press.
Savage-Rumbaugh, E.S. (1986) *Ape Language: From Conditioned Response to Symbol*, New York: Columbia University Press.
Templer, V.L. & Hampton, R.R. (2012) Rhesus monkeys (Macaca mulatta) show robust evidence for memory awareness across multiple generalization tests, *Animal Cognition*, **15** (3), pp. 409–419.
Terrace, H.S., Petitto, L.A., Sanders, R.J. & Bever, T.G. (1979) Can an ape create a sentence?, *Science*, **206** (4421), pp. 891–902.
Veit, W. (2022a) Complexity and the evolution of consciousness, *Biological Theory*, **18**, pp. 175–190.
Veit, W. (2022b) Integrating evolution into the study of animal sentience, *Animal Sentience*, **7** (32).
Veit, W. (2022c) Towards a comparative study of animal consciousness, *Biological Theory*, **17** (4), pp. 292–303.
Veit, W. (2023) *A Philosophy for the Science of Animal Consciousness*, 1st ed., London: Routledge.
Veit, W., Dewhurst, J., Dołega, K., Jones, M., Stanley, S., Frankish, K. & Dennett, D.C. (2019) The rationale of rationalization, *Behavioral and Brain Sciences*, **43**, e53.
Veit, W. & Browning, H. (2023) Social robots and the intentional stance, *Behavioral and Brain Sciences*, **46**, e47.
Veit, W., Gascoigne, S.J.L. & Salguero-Gómez, R. (2023) Evolution, complexity, and life history theory, *Authorea*, preprint. doi: 10.22541/au.167770655.56360178/v1
Veit, W. & Browning, H. (forthcoming) Neural networks, AI, and the goals of modeling, *Behavioral and Brain Sciences*.