Studies of Primate Metacognition are Relevant to Determining What Form Introspection Could Take in Different Intelligent Systems

Maisy D. Englund and Michael J. Beran

Georgia State University

Abstract

Comparative research assessing metacognition in nonhuman animals contributes to the question of what form introspection could take in humans, nonhumans, and other possibly conscious systems. We briefly review some major findings in comparative metacognition research, including some discoveries in areas looking at self-regulation and self-control. We discuss what data exist to address the three conditions for introspection defined by Kammerer and Frankish (2023) in their target article. We suggest that two of three conditions are met by existing data from nonhuman primates, and that the third condition may be more difficult, but perhaps not impossible, to assess. We argue that a comparative and developmental approach to this question of how to define and measure introspection is a productive avenue to make progress in this area.

Main Article

In their article, Kammerer and Frankish (2023) pose the hypothetical question "*What forms could introspection systems take?*" The authors emphasize the importance of doing away with narrow definitions of introspection that can only be applied to humans, allowing instead for more broad and liberal conceptualizations that could theoretically be applied to nonhuman animals and artificial intelligence systems. With this in mind, the authors choose to define introspection as *a process by which a cognitive system represents its own current mental states, in a manner that allows the information to be used for online behavioural control.* Furthermore, they present a hypothetical map of potential introspective spaces, varying along axes of directness, conceptualization, and flexibility, and five features which could characterize the content of introspective states, including direction of fit, perceptual/cognitive, intentionality, relationality, and phenomenality. They call on experts in a wide variety of fields to discuss introspection as it may or could possibly present itself in our respective subjects of study. We appreciate this call for ideas on their framework, and we are happy to provide a perspective from comparative cognitive science, and particularly primate cognition research.

Kammerer and Frankish (2023) state that there are three conditions in their definition of introspection. First, introspective processes *target* current mental states. Second, introspective processes, by their *nature*, are representational. Third, introspective processes *function* to enable a cognitive system to use information about its own mental states for online behavioral control. In thinking of research specifically focused on the possibility of metacognitive processes in nonhuman primate species, we argue that conditions one and three are met in considering that research area. We are agnostic as to condition two, for reasons we will explain below.

Animal metacognition research shows conclusively, through a variety of paradigms, that metacognitive responses in different tasks are not the result of associative learning alone and are not the result of using only public (external) information from the environment (Beran et al., 2012; Smith, 2009). Rather, at least some tests of animal metacognition afford response classes (i.e., metacognitive responses) that target current mental states alone. Some of the original tasks used with primates were susceptible to claims that what appeared to be "metacognitive" responses instead could have been the result of the animals' associations between specific stimuli present at the time of response and the previous reinforcement rates for those stimuli (Carruthers, 2008; Hampton, 2009; Jozefowiez et al., 2009; Kornell, 2014; Smith et al., 2009). For example, rhesus monkeys opted out of trials that were objectively most difficult for them when they had to classify pixel boxes on a screen as dense or sparse (Smith et al., 1997). But, with the boxes present at the time of responding, the monkeys could have relied on previously associated reward rates for "dense" or "sparse" responses to those stimuli. If so, the "opt out" response may have become conditioned as a way to avoid punishment rather than requiring any monitoring of internal states of certainty or uncertainty. These concerns, however, were dismissed in later instantiations of this task and other perceptual classification tasks and memory tasks (e.g., Basile et al., 2015; Brown et al., 2017, 2019; Call, 2010; Hampton, 2001; Hampton et al., 2004; Smith et al., 2009, 2012; Templer & Hampton, 2012). In one variation (Smith et al., 2006), monkeys had to perform blocks of trials classifying such pixelated stimuli, and then at the end of the block, they received summary feedback that was delivered *out of order* with their responses. For example, if they were correct on trial 1, incorrect on trial 2, opted out of performing trial 3, and were correct on trial 4, they received two reward pellets in a row (for trials 1 and 4), then a "timeout" for the incorrect trial (that actually was their second response), and then no feedback for the trial they opted out of (which was their third response in the block). With this approach, it was very difficult or even impossible to specifically link individual feedback (reward or timeout) to specific responses in any way that could be explained associatively. Rather, the monkeys had to have a general sense of how well they could classify a stimulus on each trial, and then use their summarized feedback to adjust their responding in subsequent trial blocks if they wanted to become more efficient. And, the resulting patterns of opting out of trials (i.e., indicating their uncertainty) showed that sometimes the objective difficulty of trials was misaligned with the use of the opt out response (i.e., monkeys opted out of stimulus densities they were actually quite good at classifying when they did classify them, and they chose to classify some densities that they actually were very bad at correctly classifying). This was a form of metacognitive illusion (Ferrigno et al., 2017; Kornell, 2014). If the target of such opt-out responses was the association of the stimuli themselves with reward or punishment histories, this misalignment could not occur. But, if the monkeys instead were relying on their mental states (perceived difficulty of the classification) as the target of their choice to classify or opt out, this pattern could emerge.

These metacognitive judgments of animals can be retrospective in nature (confidence or uncertainty about what was classified or remembered) or prospective (confidence or uncertainty about how well one will classify or remember something correctly; Morgan et al., 2014). This is again true even when the stimuli to be remembered or classified are no longer visually present and also when those stimuli may evoke different memory processes, as in the case of allowing monkeys to use familiarity versus requiring them to engage more effortful, working memory processes (Brown et al., 2019; Templer et al., 2018) In some cases, that confidence or uncertainty is reflected in wagering behavior, and such wagering can fluctuate depending on other aspects of the significance of being correct or incorrect such as how much potential reward was at stake (e.g., Kornell et al., 2007).

In addition to providing evidence of metacognitive processes that specifically target mental states, many of the above studies also included evidence of the *functional* aspect of an introspective system to enable online behavioral control depending on those differing mental states of confidence or uncertainty. To expand on one such example, Zakrzewski et al. (2014) showed that the same objective stimulus could evoke different levels of confidence depending on the stakes of making classifications. Rhesus monkeys were trained that they would receive one "banked" token for each correct classification of density boxes (i.e., dense or sparse), but reward only came when they "cashed out" their balance in that bank, and any error in classification emptied the bank of all collected tokens. "Cash out" responses could be made at the same time the current trial's stimulus was present, so that the monkeys could assess the difficulty of the present trial in the context of how much was at stake. What the monkeys showed was a greater likelihood for "cashing out" when more tokens were banked than less tokens. In other words, when few tokens were at stake, the monkeys were more likely to classify a wider range of stimuli, including many of those that objectively were very difficult. However, with more tokens at stake, a wider range of stimuli in the most difficult part of the stimulus continuum induced cash-out responses, indicating that the monkeys were monitoring the stimuli to determine how confident they were in classifying those, and monitoring the stakes in being right or wrong in that determination.

We have developed another form of confidence paradigm that can be used across species in which confidence is indicated by bodily movements through space (Beran et al., 2015). In the first use of this task, chimpanzees were given computerized memory tasks or a task in which they had to match photographs of real-world images to their associated lexigram symbols. Correct responses led to food reward, but this reward was delivered in another area, away from where task responses were made. And, importantly, those rewards were only available for retrieval if the chimpanzees were present in front of the dispenser when they were delivered a few seconds after the correct response was made on the computer. Equally important was that feedback on correctness was delayed, leaving a time window in which the chimpanzees only knew what response they made, but not whether the program had scored that as correct or incorrect. Leaving early to move to the reward dispensing location, before any feedback from the computer, was optimal in ensuring that food could be obtained and not lost, but leaving early was effortful and futile if an incorrect response had been made. Chimpanzees showed clear patterns of going to the dispenser early, before any feedback was given, more often when they would be correct than when incorrect. This pattern also was demonstrated by preschool age children (James et al., 2021) and some (but not all) capuchin monkeys (Smith et al., 2020).

We believe some of the strongest evidence of introspective states in nonhuman primates may come from tests not designed to assess metacognition specifically. In a test of self-control, chimpanzees learned that food rewards would slowly accumulate in an apparatus that they could take at any time and then eat all accumulated rewards. But, doing that stopped the delivery of even more rewards, and so the longer they waited, the more food they could obtain (Beran, 2018). Evans and Beran (2007) asked whether chimpanzees could engage in a form of self-distraction that would facilitate longer delay. In the baseline condition, food items accumulated and chimpanzees simply had to wait as long as they could. In a second condition, various toys and other items such as string or magazines were provided during the trial, and the chimpanzees could engage these things during the delay interval while food items accumulated within their reach. As expected, when such items were present, the chimpanzees used them, and delay intervals were longer. This proves nothing about self-distraction, as mere availability of things to engage may lead to engaging with those things, and this would then *externally distract* the chimpanzees from the accumulating food reward. But, the key data come from a third condition, in which the same kinds of toys and items were available, but now the accumulating reward was out of reach so that the chimpanzees were unable to control how long they had to wait to receive their reward (rewards were yoked in their time-to-delivery and amount to a trial in the second condition to ensure identical reward amount across the whole study). If the mere presence of toys and the frustration of not being able to eat visible food items was all that mattered, levels of engagement with the toys should have been identical in both conditions. However, three of four chimpanzees were more likely to engage the toys when they also had to impose on themselves the maintenance of delayed gratification (i.e., the food items could be taken at any time) compared to when that delay was externally imposed (i.e., the food was still present, but out of reach). This result suggests that the chimpanzees had some awareness of their own susceptibility to failure in waiting through the delay, and then adjusted their engagement with the toys to offset that susceptibility and wait out the delay interval. It could be that the chimpanzees were monitoring their impulsivity and adjusting behavior accordingly to other things in their environment, in much the same way that successful children employ such self-control strategies (Mischel, 2015). We know that those chimpanzees that show the highest levels of general cognitive abilities are also those individuals who are best at opting into or opting out of a delay of gratification task, suggesting again that they have some awareness of how impulsive (or self-controlled) they are likely to be on a trial-by-trial basis (Beran & Hopkins, 2018).

This leaves the second condition in Kammerer and Frankish (2023)'s definition of introspection, that introspective processes are representational. At present, and perhaps inevitably, we cannot know whether mental states are represented as such by animals. If such demonstrations require explicit declaration that an organism is in a state of representing its mental states (e.g., Carruthers, 2014), it is hard to see how such data can come from non-linguistic species. If this is a necessary condition for introspection, then it is as challenging for those who study nonhuman animal introspection and metacognition to address as the idea that only autonoesis (Tulving, 2005) is sufficient for true episodic memories for those who study nonhuman animal memory (e.g., Clayton et al., 2001; Malanowski, 2016). The compromise in that area of research has been the term "episodic-like memory." So, do we call the phenomena from metacognition studies "introspection-like" (see Melcalfe & Son, 2012)? Kammerer and Frankish do not require mental states to be represented as such, and thus the processes under consideration here can be called introspective even though they may be distinct from the processes of human introspection. Thus, introspection-like does not seem to be a necessary compromise.

Moving away from specific demonstrations of animal metacognitive abilities, we return to the question of whether animal behavior and cognition studies beyond the primates are informative about introspection more broadly. Kammerer and Frankish (2023) discuss approaching the question of introspection from the minimal mind perspective: that is, taking the most basic cognitive system (one that is able to perceive the world and respond flexibly, but lacks introspection) and consider what must be added to this minimal mind in order to evoke the ability to introspect. A comparative approach is an excellent strategy to address this question.

Comparative researchers know, with a relatively high degree of certainty, that at least some animals likely *cannot* introspect. Therefore, we can document the behavioral patterns of how a 'minimal mind' (one without introspective abilities) responds to certain situations and experimental paradigms. We also know that humans *can* introspect, and we know how humans tend to behave in similar situations and experimental paradigms. We therefore know quite a bit about what the behavior looks like on either end of this hypothetical introspection spectrum. By including more species and comparing their behavioral (and sometimes physiological or neurological) patterns to that of a known introspective mind (e.g., humans) and that of a presumably non-introspective mind (e.g., pigeons; Inman & Shettleworth, 1999; Sutton & Shettleworth, 2008; but see Adams & Santi, 2011; Templer et al., 2017), we can shed light on which other species may have some level of introspection, as well as what other cognitive faculties or evolutionary pressures may be necessary pre-requisites for such an introspective mind. Using a comparative approach (along with developmental approaches) can help us understand the building blocks of the introspective mind (Beran et al., 2012), even if we cannot conclude with certainty that any other species possess all of the conditions that are part of the definition of Kammerer and Frankish.

Animals do not have language; we cannot ask them "What prompted this response?" or "How did you think about this problem?" because they cannot answer in any way we could interpret. Even if they could, we could not know that their answers were accurate, just the same as we know that humans sometimes confabulate reasons for why they do what they do. Consequently, comparative researchers are accustomed to making inferences about the mechanisms underlying animals' behaviors by ruling out as many alternative explanations as possible, and ultimately assuming that the simplest (remaining) explanation is the most likely to account for the results. Therefore, with regards to Kammerer and Frankish's (2023) mapping of possible introspective devices, comparative researchers would start by assuming that, at its most basic level, introspection would be *direct* and *non-conceptual.* That is, if animals do possess a level of introspection, we have no reason to believe that there are several degrees of separation between first-order mental states (e.g., experiencing a state of anxiety or unease) and the ultimate introspective representation (e.g., a known feeling of uncertainty about a situation). Nor do we have any reason to believe that the 'outputs' of such introspection would more likely be conceptual (akin to beliefs) than to be non-conceptual (akin to sensations). The most parsimonious explanation for a basic introspection device, as it fits the definition outlined by the authors, would be primarily an analogue process, whereby the first-order cognitive state directly causes the introspective state, evoking a (non-conceptual) representation of the current mental state (the monitoring aspect of metacognition), which may then be accessed to create a behavioral response (the control aspect of metacognition)(Flavell, 1979). There is evidence, albeit limited, that nonhuman animals engage in some of the same forms of abstract thought or conceptualization that humans can engage. Despite this, there is no reason to believe that those higher-order processes are more likely to be the target of introspection than lower-order processes of which we know animals are capable. Therefore, if we assume that animals have some level of introspection that is likely to be non-conceptual, their introspective states are likely to be of a more direct and non-conceptual nature, unless there is evidence that rules out this parsimonious explanation. And, even before we can do that, we must first rule out that any behavior that looks like introspection can be explained by another simpler explanation, such as associative learning. As noted earlier, this was the challenge to the preliminary reports about uncertainly monitoring in the animal metacognition literature, and that challenge was met. So, it is possible that creative paradigms can emerge to address this same concern about other introspective experiences.

Studying introspection through a comparative lens can help identify its proximate (i.e., mechanistic) and ultimate (i.e., evolutionary) causes. By studying multiple species and their respective cognitive abilities, we can help identify which other cognitive faculties are necessary for, or at least correlated with, the ability to introspect. Furthermore, we can consider what selective pressures would lead to the ability to introspect and make testable predictions based on these theories. In short, comparative research can shed light on the *function* of introspection in a way that other approaches may not.

Kammerer and Frankish (2023) state the following: “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.” We agree wholeheartedly with this perspective. We have advocated that research focused on the metacognitive abilities of animals lends itself to this better perspective. It may be, and probably is, the case that some aspects of adult human metacognition are unique to humans, but the goal of comparative research is not to classify all human cognitive faculties as unique or fully shared, and then stop. Rather, the goal should be to see many such faculties as being sufficiently complex and adaptive that even “minimal versions” as evidenced in other species are informative about our own species, and about the evolutionary pressures and pathways that allowed such faculties to become more advanced over time. As with the first primitive eyes, primitive feathers, or primitive communicative signals evidenced by the ancestors of species that now have excellent vision, flight, or sophisticated communication systems, there can be forms of metacognition and introspection in other species that illustrate the path to *Homo sapiens’* eventual demonstration of the strongest manifestations of those capacities.

References

Adams, A. & Santi, A. (2011) Pigeons exhibit higher accuracy for chosen memory tests than for forced memory tests in duration matching-to-sample, *Learning & Behavior,* **39***,* pp. 1–11.

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. J. (2018) *Self-control in Animals and People,* London: Academic Press.

Beran, M. J., Brandl, J., Perner, J., & Proust, J. (eds.) (2012) *Foundations of Metacognition,* Oxford: Oxford University Press.

Beran, M. J. & Hopkins, W. D. (2018) Self-control in chimpanzees relates to general intelligence, *Current Biology* **28***,* pp. 574-579.

Beran, M. J., Perdue, B. M., Futch, S. E., Smith, J. D., Evans, T. A. & Parrish, A. E. (2015) Go when you know: Chimpanzees' confidence movements reflect their responses in a computerized memory task, *Cognition* **142***,* pp. 236-246.

Brown, E. K., Basile, B. M., Templer, V. L. & Hampton, R. R. (2019) Dissociation of memory signals for metamemory in rhesus monkeys (*Macaca mulatta*), *Animal Cognition* **22**, pp. 331-341.

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.

Call, J. (2010) Do apes know that they could be wrong?, *Animal Cognition* **13***,* pp. 689–700.

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

Carruthers, P. (2014) Two concepts of metacognition, *Journal of Comparative Psychology* **123**, , pp. 138–139.

Clayton, N. S., Griffiths, D. P., Emery, N. J. & Dickinson, A. (2001) Elements of episodic–like memory in animals, *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **356**, pp. 1483-1491.

Evans, T. A. & Beran, M. J. (2007) Chimpanzees use self-distraction to cope with impulsivity, *Biology Letters* **3***,* pp. 599-602.

Ferrigno, S., Kornell, N. & Cantlon, J. F. (2017) A metacognitive illusion in monkeys. *Proceedings of the Royal Society B: Biological Sciences* **284**, pp. 20171541.

Flavell, J. H. (1979) Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry, *American Psychologist* **34***,* pp. 906–911.

Hampton, R. R. (2009) Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms?, *Comparative Cognition and Behavior Reviews* **4***,* pp. 17–28.

Hampton, R. R. (2001) Rhesus monkeys know when they remember, *Proceedings of the National Academy of Sciences* **98***,* pp. 5359–5362.

Hampton, R. R., Zivin, A., & Murray, E. A. (2004). Rhesus monkeys (*Macaca mulatta*) discriminate between knowing and not knowing and collect information as needed before acting, *Animal Cognition* **7**, pp. 239-246.

Inman, A. & Shettleworth, S. J. (1999) Detecting metamemory in nonverbal subjects: A test with pigeons, *Journal of Experimental Psychology: Animal Behavior Processes* **25***,* pp. 389–395.

James, B. T., Parrish, A. E, Guild, A. S., Creamer, C., Kelly, V., Perdue, B. M., Kelly, A. J. & Beran, M. J. (2021) Go if you know: Preschool children's movements reflect their metacognitive monitoring, *Cognitive Development* **57***,* 101001.

Jozefowiez, J., Staddon, J. E. R. & Cerutti, D. T. (2009) Metacognition in animals: How do we know that they know?, *Comparative Brain and Behavior Reviews* **4***,* pp. 29–39.

Kornell, N. (2014) Where is the "meta" in animal metacognition?, *Journal of Comparative Psychology* **128,** pp. 143-149.

Kornell, N., Son, L. K.& Terrace, H. S. (2007) Transfer of metacognitive skills and hint seeking in monkeys, *Psychological Science* **18***,* pp. 64–71.

Malanowski, S. (2016). Is episodic memory uniquely human? Evaluating the episodic-like memory research program, *Synthese* **193**, pp. 1433-1455.

Metcalfe, J. & Son, L. K. (2012) Anoetic, noetic and autonoetic metacognition, in M. Beran, J. R. Brandl, J. Perner & J. Proust (eds.) *The Foundations of Metacognition,* Oxford: Oxford University Press.

Mischel, W. (2015) *The marshmallow test: Why self-control is the engine of success*. Little, Brown.

Morgan, G., Kornell, N., Kornblum, T. & Terrace, H. S. (2014) Retrospective and prospective metacognitive judgments in rhesus macaques (*Macaca mulatta*), *Animal Cognition* **17***,* pp. 249–257.

Smith, J. D. (2009) The study of animal metacognition, *Trends in Cognitive Sciences* **13***,* pp. 389–396.

Smith, J. D., Shields, W. E., Schull, J. & Washburn, D. A. (1997) The uncertain response in humans and animals, *Cognition* **62***,* pp. 75–97.

Smith, J. D., Beran, M. J., Couchman, J. J., Coutinho, M. V. C. & Boomer, J. (2009) Animal metacognition: Problems and prospects, *Comparative Cognition and Behavior Reviews* **4***,* pp. 40–53.

Smith, J. D., Beran, M. J., Redford, J. S. & Washburn, D. A. (2006) Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring, *Journal of Experimental Psychology: General* **135***,* pp. 282-297.

Smith, J. D., Couchman, J. J. & Beran, M. J. (2012) The highs and lows of theoretical interpretation in animal metacognition research, *Philosophical Transactions of the Royal Society B* **367***,* pp. 1297–1309.

Smith, T. R., Parrish, A. E., Creamer, C., Rossettie, M. & Beran, M. J. (2020) Capuchin monkeys (sometimes) go when they know: Confidence movements in *Sapajus apella, Cognition* **199***,* 104237.

Sutton, J. E. & Shettleworth, S. J. (2008) Memory without awareness: Pigeons do not show metamemory in delayed matching to sample, *Journal of Experimental Psychology: Animal Behavior Processes* **34**, pp. 266–282.

Templer, V. L., Brown, E. K. & Hampton, R. R. (2018) Rhesus monkeys metacognitively monitor memories of the order of events, *Scientific Reports* **8**, pp. 11541.

Templer, V. L. & Hampton, R. R. (2012) Rhesus monkeys (*Macaca mulatta*) show robust evidence for memory awareness across multiple generalization tests, *Animal Cognition* **15***,* pp. 409–419.

Templer, V. L., Lee, K. A. & Preston, A. J. (2017) Rats know when they remember: Transfer of metacognitive responding across odor-based delayed match-to-sample tests, *Animal Cognition* **20**, pp. 891-906.

Tulving, E. (2005) Episodic memory and autonoesis: Uniquely human?, In H. S. Terrace & J. Metcalfe (eds.), *The Missing Link in Cognition: Origins of Self-Reflective Consciousness,* Oxford: Oxford University Press.

Zakrzewski, A. C., Perdue, B. M., Beran, M. J., Church, B. A. & Smith, J. D. (2014) Cashing out: The decisional flexibility of uncertainty responses in rhesus macaques (*Macaca mulatta*) and humans (*Homo sapiens*), *Journal of Experimental Psychology: Animal Learning and Cognition,* **40***,* pp. 490–501.