Self-Consciousness in Animals and Artificial Agents

Michael J. Hegarty

September 30th, 2025



Part 1: Self-Consciousness in Goffin's Cockatoos

Background: Summary of Claims

- (See Hegarty (2025 & MS).
- Goffin's cockatoos are capable of a kind of creative, 'individual innovation' in tool use.
- To be capable of this, it is likely that they are self-conscious in the minimal sense of being capable of *nonconceptual self-representation*.
- The function of nonconceptual self-representation is to unify different representations—specifically, those anchored to the here-and-now with *prospective* representations (about non-actual states of affairs).

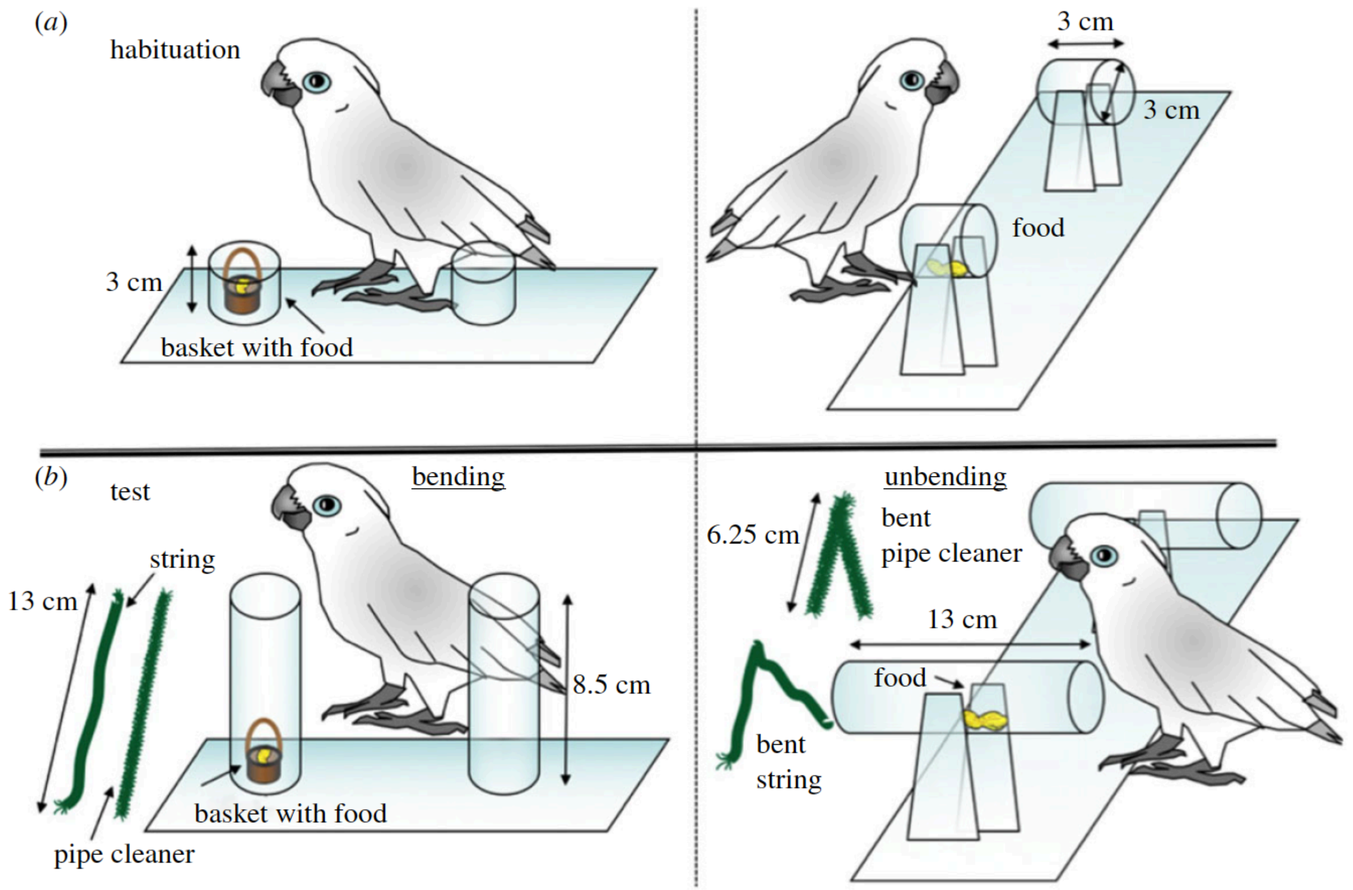


Bridging Previous Work & Machine Learning



- Cognitive neuroscientists think that the sort of learning and problem-solving I describe in the cockatoos is modelled by a certain class of machine learning algorithms: temporal-difference (TD) learning algorithms.
- So, if I am right that a minimal kind of self-consciousness functions to unify temporally distinct representations, then would an artificial intelligence that solves analogous problems by employing a TD algorithm also be self-conscious?

Empirical Case Study of Tool Use: Laumer et al. 2017



rspb.royalsocietypublishing.org Proc. R. Soc. B 284: 20171026

Figure 1. (*a*) Apparatuses for the habituation phase; (*b*) apparatuses for the testing phase. Left: vertical tube apparatus; right: horizontal tube apparatus. Dimension are given in centimetres.

Laumer et al. 2017

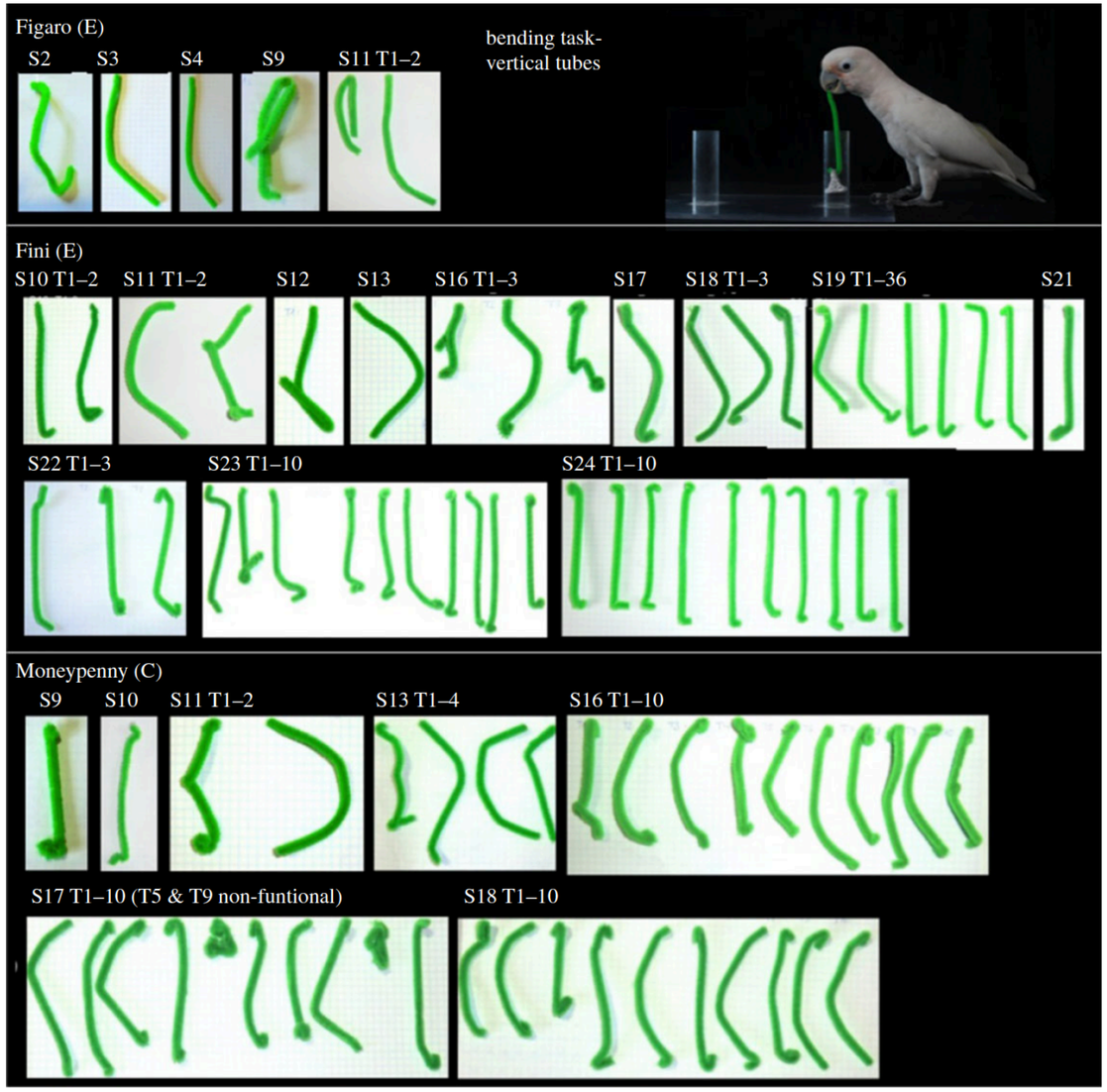


Figure 3. Successful hook-tools build by subjects Figaro, Fini and Moneypenny in the vertical tube condition. S, session number; T, tool number; (C), control group/ no experience; (E), experience group; Photograph by Solvin Zankl.

rspb.royalsocietypublishing.org Proc. R. Soc. B 284: 20171026

Laumer et al. 2017

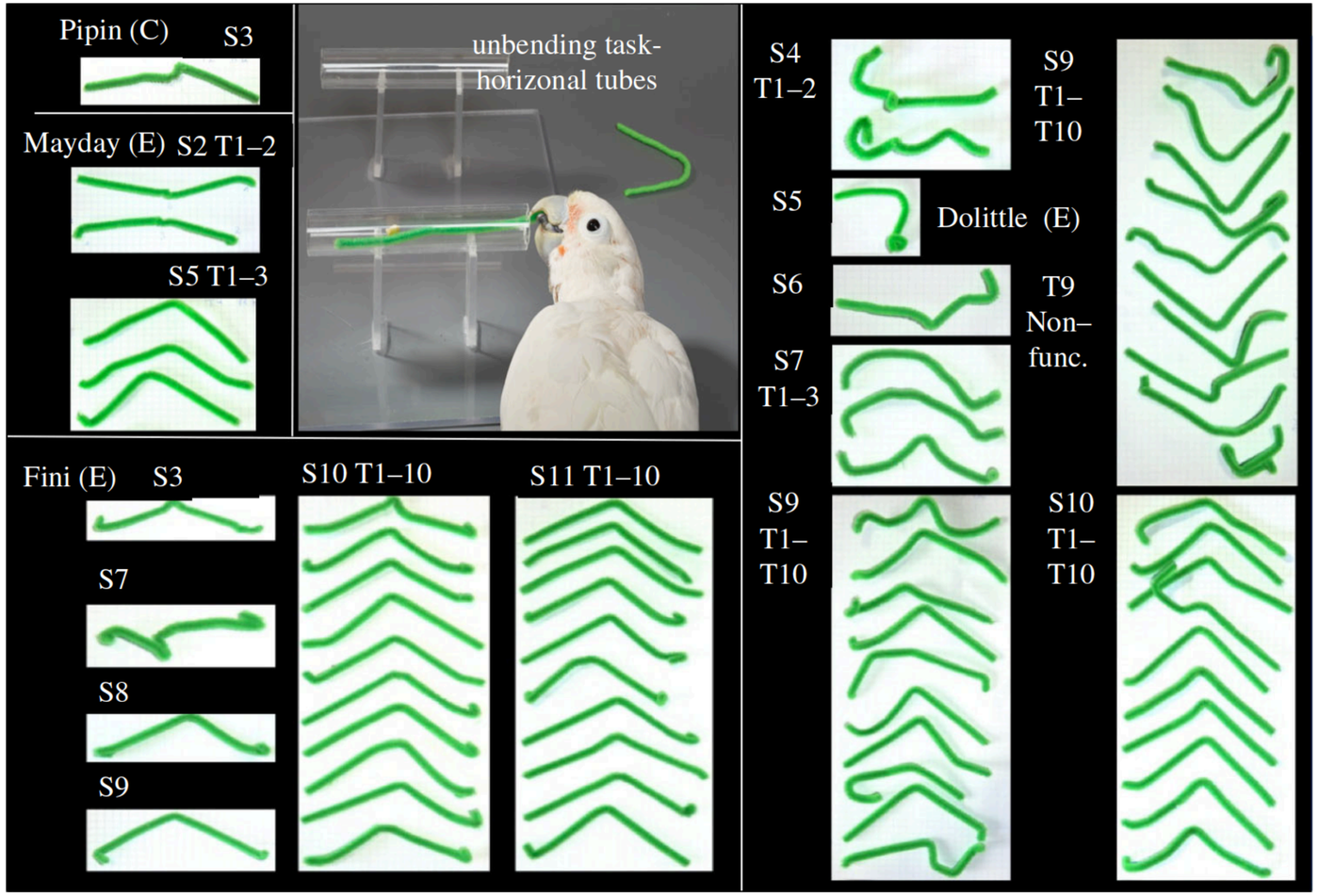


Figure 4. Successful stick-tools build by subjects Pipin, Mayday, Fini and Dolittle in the horizontal tube condition. S, session number; T, tool number; (C), control group/no experience; (E), experience group; Photograph by Bene Croy.

rspb.royalsocietypublishing.org Proc. R. Soc. B 284: 20171026

Instrumental Practical Rationality from Individual Innovation

- Camp & Shupe: IR is to “act to achieve a goal, G, by identifying a state of affairs M, which is neither inherently desirable nor currently actual, as to-be-done because it will centrally contribute to actualizing G” (2017, p. 100).
- An individual innovator “recognizes a way to achieve a goal that it cannot achieve directly, by bringing about a subsidiary state of affairs. [T]o do this, *it must represent that subsidiary state while realizing both that this state does not actually obtain and that its obtaining would help to achieve the primary goal*… [and] the creature’s recognition of the connection between the two states must be established through spontaneous ‘insight’” (Camp 2009, p. 292, emphasis mine).

Summary of Hegarty (2025 & MS)

- If proper function cannot explain the cockatoo behaviour, then we must explain:
    1. the source of normativity for this intentional action, and
    2. what grounds attribution of cognitive states to the same subject.
- I argue by analogy with Anscombe's work on agency, and the role of 'I'-Thoughts in action explanation that we explain both by positing self-representational content common to the cockatoos' representations.
- Representations (reminder): perceptual and prospective (the non-actual state of affairs of *the wire's being bent thus-and-so*).
- Suggestion: understand both in action-centric, practical, and embodied terms.



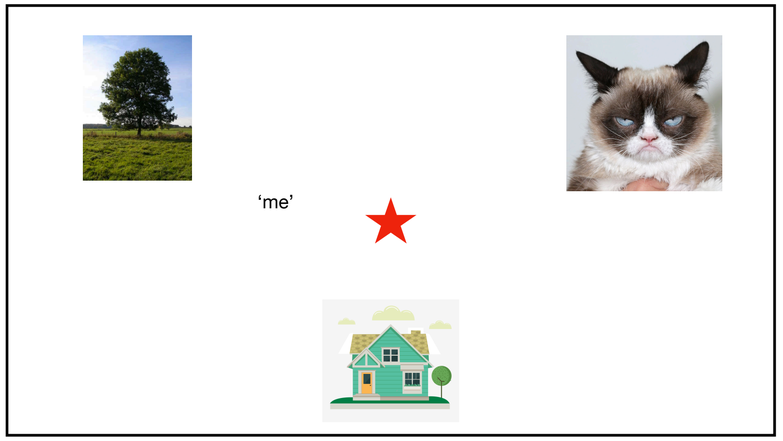
- Madman! The ‘brutes’ aren’t that sophisticated. Surely cockatoos have no self-conception, and can’t think ‘I’-thoughts.
- But: drawing on nonconceptual self-consciousness (e.g. in somatic proprioception, perception of Gibsonian affordances) we dodge this worry (c.f. Bermudez 1998, Musholt 2015).
- Nonconceptual self-representation ≠ capacity to think ‘I’-Thoughts.

Egocentric versus Self-Representational Cognitive Maps

Egocentric cognitive map (no self-representation)



Self-representational cognitive map

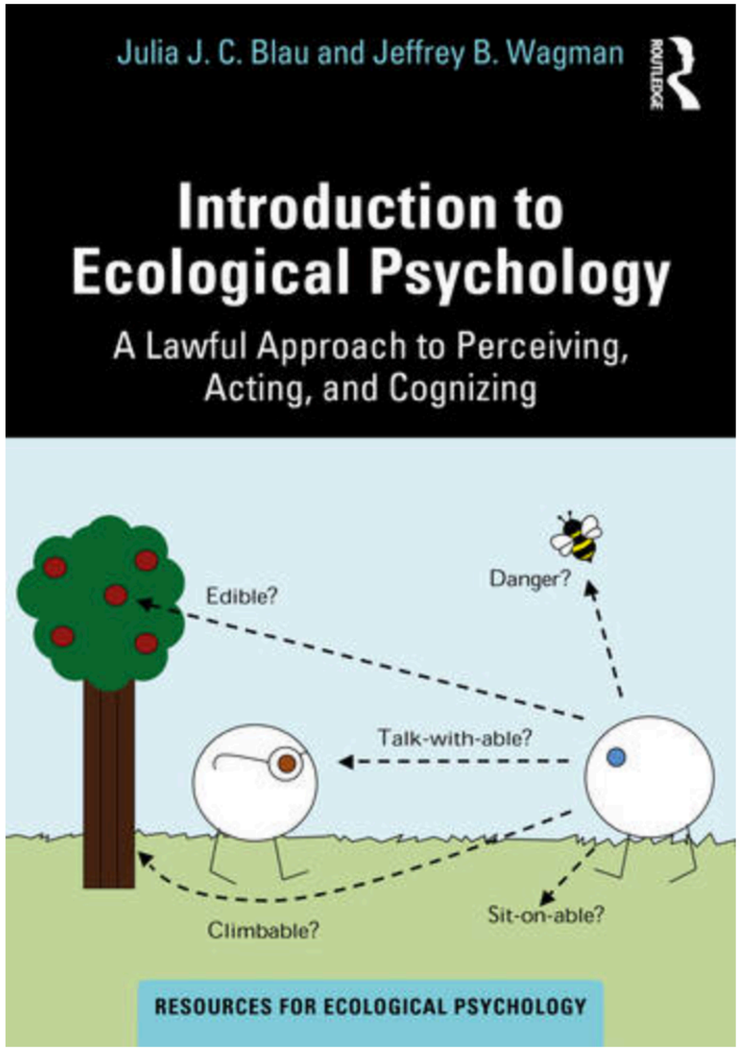


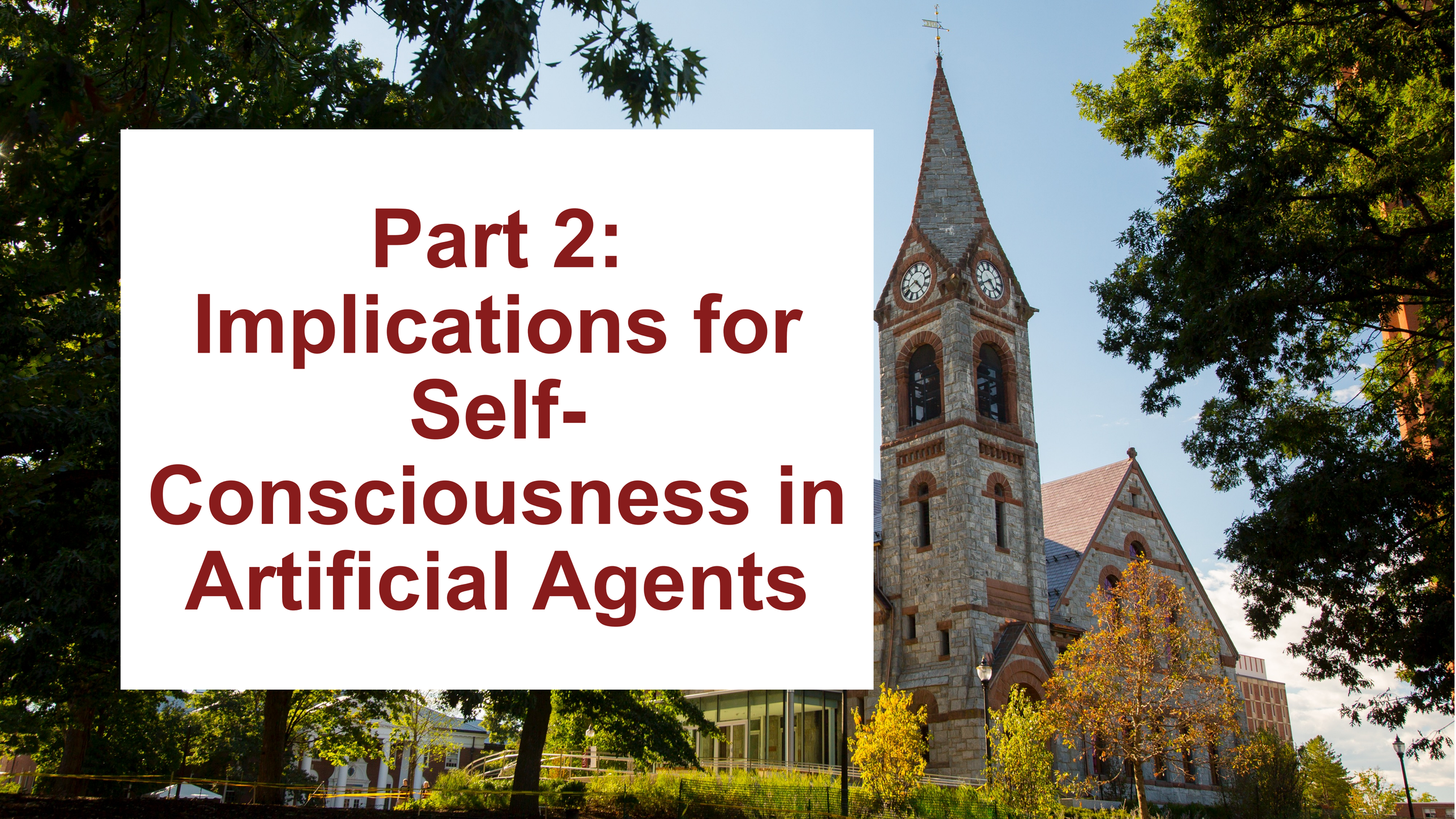
Content: ⟨to-be-eaten, __ in front of/behind/etc. [here]⟩ **Content: ⟨to-be-eaten, __ in front of/behind/etc. [me]⟩**

Rough difference: information entered into a 'here' file versus information entered into a 'me' file. To even have a 'me' file requires some idea of a 'moveable here' with identity conditions and continuity through time

Egocentric Cognitive Maps & Affordance Perception

- I propose to treat perceptual and spatial representations as involving ‘affordances’, or possibilities for action (c.f. Millikan 2004).
- Hegarty (2025) argues that the representation of the non-actual state of affairs, $\Re\not\exists$, cannot appear on a cognitive map.
- But suppose that animals with affordance perception can ‘follow’ those affordances in thought through a kind of embodied imagining.
- So, $\Re\not\exists$ may be possible for a cognitively modest creature.





Part 2: Implications for Self-Consciousness in Artificial Agents

Reinforcement Learning

- Reinforcement Learning (RL) refers both to (i) a kind of learning problem, and (ii) a family of machine learning approaches (algorithms) for solving that problem.
- Roughly: trial-and-error learning of how to act to maximise reward through exploratory interactions with an incompletely understood environment.

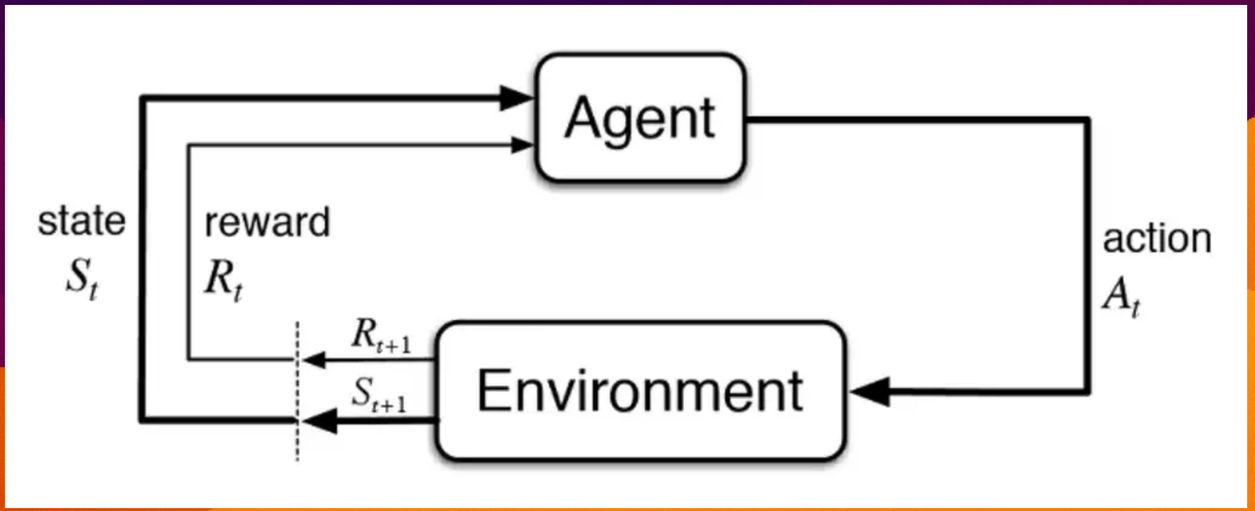
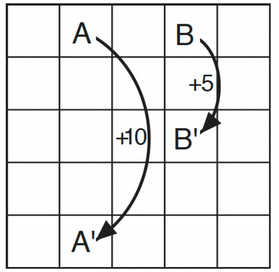
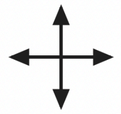


Diagram from Sutton & Barto (2018)





Actions

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

A Question

- Cognitive neuroscience orthodoxy: some brain processes correlated with animals solving RL problems appear to be implementing RL algorithms (Dayan 2009; Momennejad et al. 2017).
- The cockatoos are facing an RL problem: how to get food by exploring their environment and exploiting its features.
- So, one might wonder…
- If (i) the cockatoos solve RL problems because RL algorithms are biologically implemented in them, ***and*** if (ii) nonconceptual self-representation is required for their cognition, ***then*** are artificial intelligences/agents using RL algorithms to solve RL problems *also self-conscious*?

Are RL-implementing AIs Self-Conscious?

- My answer will be a qualified ‘no’.
- My argument for self-consciousness in the cockatoos is at a different level of analysis from the claim about RL algorithm implementation.
- Moreover, the computational problem (i.e. functional description of the cognitive task to be performed) I invoke self-consciousness to solve is different—by being essentially embodied—from the computational problem the RL algorithm solves.

Levels of Analysis in Cognition

1	Computation	What the cognitive task is functionally, i.e., in terms of its input and output
2	Algorithmic	How the task is performed, i.e., how the output is calculated for each input
3	Implementation	How the algorithm is implemented on the physical level

Marr's Levels. Diagram from Pantsar (2021)

- Standard distinction in philosophy of mind/cognitive science: various levels of analysis at which mental and cognitive phenomena can be described, and various levels at which explanations of those phenomena operate (Marr 1982; Pylysyhn 1984; Houng 2012).

The Dopamine Reward System

- At what level is the neuroscientist's claim that RL algorithms (strictly: temporal-difference (TD) algorithms) are implemented?
- Neuroscientists have investigated "the relationship between reinforcement learning algorithms and reinforcement learning in the nervous system" and have found an "uncanny similarity between the behavior of temporal-difference algorithms and the activity of dopamine producing neurons in the brain" (Sutton & Barto, 2018, 21).
- So: the claim is at the implementational/ algorithmic level.

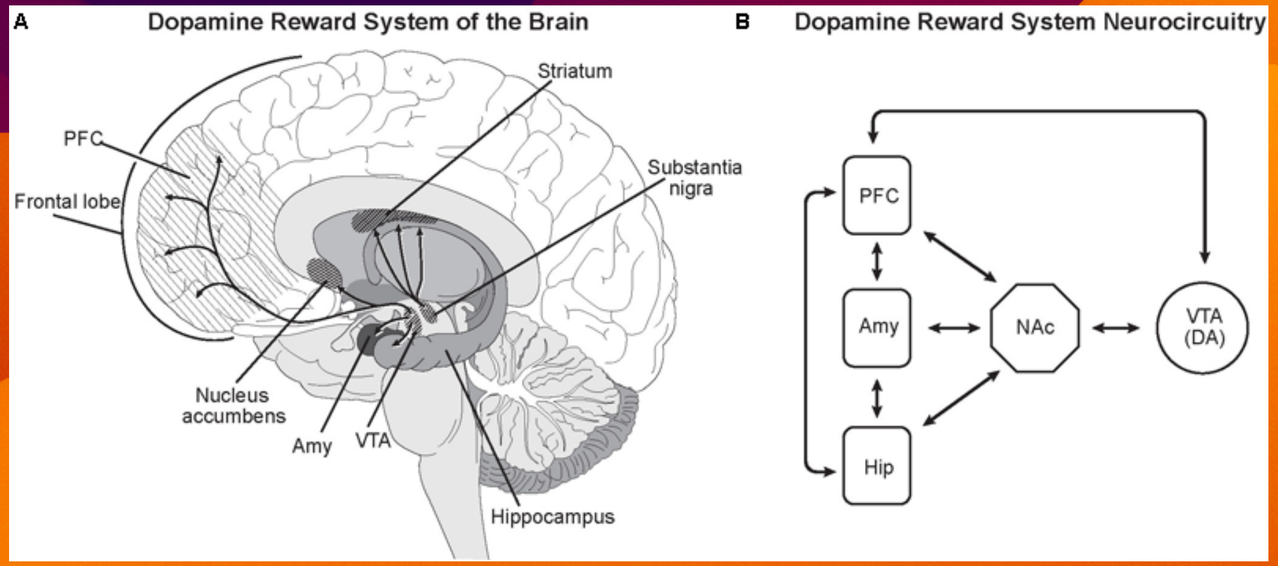


Diagram from Sonne & Gash (2018)

Levels of Analysis in Cognition II

1	Computation	What the cognitive task is functionally, i.e., in terms of its input and output
2	Algorithmic	How the task is performed, i.e., how the output is calculated for each input
3	Implementation	How the algorithm is implemented on the physical level

Marr's Levels. Diagram from Pantsar (2021)

- Cognitive psychologists typically work top-down, from a psychological-functional level of description.
- Cognitive neuroscientists typically work bottom-up, from an algorithmic-neurobiological/implementational level of description.
- My argument has nothing to do with neurological implementation, or the particular algorithmic implementation of functionally described tasks.
- My work is closer to cognitive psychology than cognitive neuroscience.
- (See: Datteri & Laudisa (2014) and Kriegeskorte & Douglas (2018) for scientific perspectives on relating psychology and neuroscience.)

Inferences Between Levels?

- There is no robust, widely-held theory of how to relate different levels of analysis.
- So, inferences across levels are not generally valid.
- And the question posed depends on such an inference.

A Better Answer?

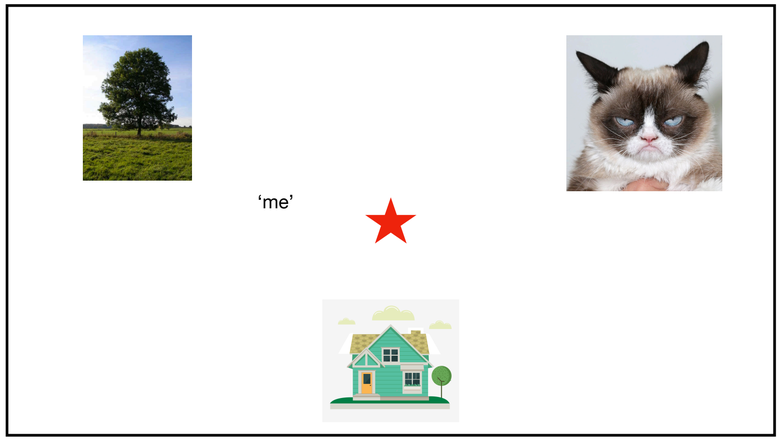
- I think that's a little unsatisfying. A more convincing reply might engage with the substance of the functional tasks that nonconceptual self-representation and the RL algorithms are performing.
- But carefully distinguishing levels of analysis also puts us in a position to say something along those lines.
- Nonconceptual self-representation functions to combine prospective and here-and-now anchored representations. But an RL/TD algorithm functions to maximise a numerical reward signal.
- Divergent functional tasks → not the case that any implementation of a TD algorithm implies self-representation.

Egocentric versus Self-Representational Cognitive Maps

Egocentric cognitive map (no self-representation)



Self-representational cognitive map



Content: ⟨to-be-eaten, __ in front of/behind/etc. [here]⟩ **Content: ⟨to-be-eaten, __ in front of/behind/etc. [me]⟩**

Rough difference: information entered into a 'here' file versus information entered into a 'me' file. To even have a 'me' file requires some idea of a 'moveable here' with identity conditions and continuity through time

If Not This, Then What? A Potential Representational Explanation

- Assuming a representational theory of mind, then the following representations would explain its behaviour (complete discussion in Hegarty (2025 & ms)):

1. Goal state: desire for the food in cylinder.

2. Perceptual states: sees the food in the cylinder, and sees the pipe cleaner.

3. $\Re\not\exists$: representation of the non-obtaining state of affairs *the pipe cleaner's being bent thus-and-so*.

A Better Answer?

- The representational content to be integrated depends on embodiment.
- This is dictated by the shape of the problem the cockatoos face, and also supports a non-over-intellectualised, action-centric account of their cognition.
- The TD algorithm describes the flow of information during computation. It solves the computational problem of how to maximise expected reward.
- Self-representation solves a different computational problem: how to integrate states representing non-actual states of affairs (embodied imaginings arrived at by mentally following perceived affordances) with perceptual and bodily states anchored to the here-and-now (where these states have egocentric content).
- The TD algorithm works just fine without indexing any states to ‘me’ or ‘this body’. But I argued that the cognitive function in the cockatoos (probably) requires common self-representational content (‘self’-indexing).

Conclusion

- Putting things perhaps too simply: the TD algorithm solves an abstract learning problem, whereas the learning problem the cockatoos face is not aptly thought of as an abstract learning problem that happens to be implemented in bodies.
- TD might be involved, at a low-level, in how the cockatoos cognise. But, if there's no analogue to the embodied combination problem, I see no reason to think TD implementation alone suffices for self-representation.