PRE-PRINT DRAFT 1

Meta-management as a clearer explanation of conscious function

Malcolm J. Lett

Independent.

Author Note

ORCiD: https://orcid.org/0000-0003-4903-1580

Email: malcolm.lett@gmail.com

Website: malcolmlett@github.com

Abstract

Conscious awareness is explained as a result of meta-management processes in the brain which are required in order to control cognitive state-space trajectories during deliberation. It is reviewed in the context of evolution and the need to cope with exponential increases in environmental and social complexity while avoiding exponential increases in brain size. A mechanism for auto-meta-management is explained, via a cognitive-state feedback loop made available as a first-class sense. The result forms the contents of consciousness. Three phenomenological aspects of consciousness are given clear explanations: that it is limited in scope and detail, that it "looks through" to first-order states, and the timeliness of conscious awareness.

Keywords: Meta-management, consciousness, conscious awareness, conscious contents, access consciousness, intentionality, learning, evolution.

Meta-management as a clearer description of conscious function

Theories of consciousness seek to explain various aspects of the fact that as humans we have subjective first-person experiences of various things such as our sensory perceptions (Gleitman, 2004, p. 204-237), our goals (Smithies and Weiss, 2019; Schroeder, 2020), and our thoughts (Gleitman, 2004, p. 278-315). Many theories have been proposed, but few are able to clearly articulate *why* consciousness evolved, *how* its underlying mechanisms function in order to produce the particular characteristics of conscious experience, and *what* makes conscious experience so special. One promising line of investigation involves *meta-management* (Sloman, 1988, 2008; Beaudoin, 1994). Here, separate first-order and second-order control processes operate in tandem. The first-order process is concerned with the major coordination of the individual's physical form in order to respond to the environment and to meet the individual's needs. The second-order process monitors and controls the first-order process at a higher level of abstraction, detecting when the first-order process makes mistakes and providing part of the training system for that first-order process. However, there are still significant gaps in our understanding of how meta-management might function in the brain and how exactly it relates to conscious experience.

This paper offers an explanation of consciousness in terms of meta-management, by first explaining the evolutionary problem that meta-management solves, by suggesting a foundational computational structure that supports meta-management, and finally by illustrating how such a structure produces a number of phenomenological characteristics associated with conscious experience. In so doing, it provides a unifying explanation for many puzzles of consciousness and offers insights for better understanding human intelligence.

Background

Researchers taking inspiration from human-level intelligence in order to build artificial intelligence systems have noted that complex multi-task deliberative systems likely require some form of meta-management. For example, it may be required in order to identify the most effective strategies among a repertoire of possible strategies for a given scenario (Sloman, 1998); to aid in the selection, orchestration, and training of separate "modules" devoted to certain skills (Sloman, 2008); to act in support of or opposition to instinct level "alarms" (eg: emotions) that may at times be counterproductive (Sloman, 2008). Some common problems have been identified that can occur with long spells of deliberative processing, including: *oscillations* between decisions, insistent *goal disruption*, excessive *multi-tasking*, *digressions* that lose track of the original problem, and *maundering* upon a small detail without reaching a conclusion (Beaudoin, 1994).

It has been noted that the development of artificial systems with such meta-management processes could well lead to robots concluding that they are conscious (Sloman, 1998).

The majority of research into *meta* processes within the brain are discussed under the umbrella term *meta-cognition*. Many behavioural studies and theories look at meta-cognition in terms of high-level executive functions such as meta-learning (being able to adapt behaviour in order to improve learning outcomes), meta-knowledge (judgements about the level of certainty in a point of knowledge or of a memory), and meta-planning (making long-term life-planning decisions) (Fleming, Dolan, & Frith, 2012; Winkielman, Schooler, 2012). Some works look at lower-level functions, such as judgment of certainty and errors during decision making (Fernandez Cruz, Arango-Muoz, & Volz 2016; Benjamin, Bjork, & Schwartz, 1998), trading off

between exploration and exploitation (Marković, Goschke, & Kiebel, 2021), or for running statistics across the brain in order to model baseline signal to noise ratios (Lau, 2007).

The theory of Representational Redescription (RR) makes the case that meta-cognitive processes require that the brain learns to construct meta-representations - high level abstractions of knowledge - in order to support learning and ongoing judgments of certainty (Karmiloff-Smith, 1992; Clark & Karmiloff-Smith, 1993; Cleeramans et al 2007; Pasquali et al, 2010; Timmermans et al, 2012). The Radical Plasticity Theory (RPT) extends the idea by suggesting that the brain also learns meta-representations of its own state and behaviours, in order to support judgments of certainty, and monitoring and prediction of behaviours (Cleeremans, 2007, 2019; Cleeremans et al, 2020). Simulations of four possible representational redescription implementations have been examined and compared against human behavioural data: i) a single-channel model where a single first-order network produces both its prediction and a certainty measure, ii) a dual-channel model where independent pathways compute the prediction and the certainty measure, iii) a hierarchical model where a second network examines the prediction made by the first-order network (or possibly some internal state of the first-order network) in order to determine the certainty measure, and iv) a hybrid of hierarchical and dual-channel models that uses both the first-order output or state and its original inputs. The hierarchical and hybrid models have been found to best mirror human accuracy and mistakes in judgment of certainty (Cleeremans et al, 2007; Pasquali et al, 2010; Timmermans et al, 2012).

The above investigations are couched in the context of high level intelligence and general meta-cognition. What if we wanted to build a practical application of these into modern AI? These meta-cognitive examples require a wider cognitive framework in order to operate;

something which we do not yet understand. What is missing is a fundamental theory of the more low-level aspects of intelligence.

The sections that follow attempt to offer such a theory in order that I can say something insightful about conscious experience. Here I shall use the *meta-management* term to refer to those low-level processes, in order to distinguish from the higher-level meta-cognitive behaviours.

From Reaction to Rumination

The earliest biological neural networks produced simple reactionary coordination and sensory interpretation for the purpose of immediate reaction (Paulin & Cahill-Lane, 2019; Godfrey-Smith, 2016, p27-41). This is mirrored in classical and contemporary connectionist AI methods which typically produce an immediate response for every sensory "time step" (Schmidhuber, 2015; Lazaridis, 2020).

The Cambrian explosion, about 542 to 485 million years ago, saw a rapid increase in the variety of animal forms and in brain size and complexity. One likely explanation is that a change from scavenging to predation led to both predator and prey entering into an arms race of intelligence: the prey evolving ever better evading strategies and better inference of predator behaviour, and the predators evolving ever better attacking strategies and better inference of prey behaviour (Godfrey-Smith, 2016, p27-41). In social species an additional level of complexity arises in needing to understand the state of other's minds. Appropriate responses to a situation need to be mediated by an inference of the other's character and mood. Likewise, possible reactions from different people need to be considered when attempting to influence a person or group for a particular outcome.

As the environmental and behavioural repertoire becomes ever more complex, one way for the brain to evolve is to simply add more neurons to the existing sense-inference-reaction cycle; for example, by making the network wider, deeper, or both. At some point that approach is no longer sustainable, as the exponential increase in environmental and behavioural complexity would require an exponential increase in brain size. An alternative solution is to introduce a new form of brain complexity, by enabling multiple iterations of inference before producing a reaction (van Bergen & Kriegeskorte, 2020; Spoerer et al, 2020). In other words, *deliberation*. Van Bergen & Kriegeskorte (2020) make the case that recurrency is employed in biology for this very reason, and that it enables organisms to dynamically trade-off between speed and accuracy.

Such recurrency might be in the form of the kinds of explicit processing loops used in programming languages where the final output of one iteration of processing is used as the new input to the next iteration, or something more biologically likely such as the multiple sequences of forward and backward propagation suggested by *predictive coding* (Friston, 2010; Clark, 2013 and 2019; Kilner, Friston, Frith, 2007). For the purposes of this paper, deliberative processing, or simply deliberation, is considered to be any situation where a bodily action in response to a given sensory signal is produced after an *arbitrary* and *variable* period of repeated re-application of inference processes, and where during that period the brain state produced as a result of earlier inferences modulate later inferences.

Deliberation encounters a problem - that the trajectory of brain state during extended inference over time can become dissociated from the immediate needs of the organism. In other words, the bodily and environmental needs should *ground* cognitive processes, but during deliberation that grounding can be lost. The problem is that reinforcement feedback provided by the environment offers only *sparse* feedback in relation to that trajectory (Sutton & Barto, 2018).

Such feedback is only in relation to the outcome of that processing. Simply preferring shorter deliberation times is too simplistic to account for the variability in the forms of deliberation that may be required.

A visual metaphor helps to illustrate this point. Figure 1 illustrates how the physical environment includes areas that cannot be accessed via straight lines, and includes obstacles that must be avoided. Likewise, navigation within brain state-space may involve taking trajectories that are sometimes quite long in order to work through a problem, while avoiding cognitive obstacles of digressions, maundering, goal disruption, excessive busyness, and others.

Thus, evolution had to find a solution to *meta-management*: a need for second-order control over the state-space behaviours of the first-order sense-inference-reaction process.

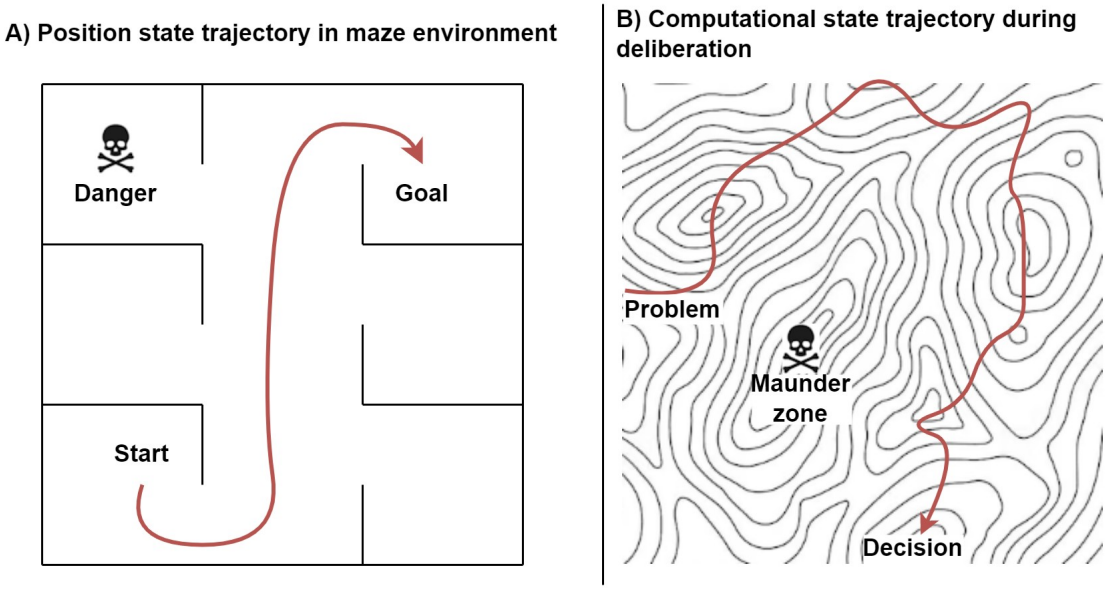

Figure 1 - Trajectories in physical and computational state-spaces. *A) Within a maze environment, the agent needs to navigate as it moves its body from start to goal. The path taken is subject to a number of constraints, such as not being able to walk through walls, and avoiding dangers. B) While deliberating, the agent moves through a series of computational states from*

the identified problem to a decision. The trajectory that it takes determines the time needed to reach the decision, and the effectiveness of the decision. Even in computational state-space some areas must be avoided, here indicated as a zone that tends to lead to maundering.

The Meta-management Process

The brain requires a few features in order to successfully carry out meta-management. Firstly, in order to control something, you must observe it. Thus, the meta-management process must observe first-order cognitive state. It cannot observe all state of the brain because that would lead to an infinite regress on the number of neurons needed to interpret and respond to that state. Thus it must observe the smallest portion of brain state, to the least level of dimensionality necessary, in order to sufficiently perform the function of meta-management. Secondly, the meta-management process must interpret the brain state and its trajectory, to draw inferences about whether the current trajectory is likely to lead to the desired outcome, and to determine what kinds of remediation might be needed. This requires that the meta-management process *models* cognitive behaviours, and that it has some understanding of the various situational *domains* within which the first-order process produces its behaviour. Lastly, the meta-management process must be able to influence the trajectory of the first-order process in some way.

There are many different ways in which such a process could be implemented within the brain. One such approach will be described here that closely resembles what we have come to understand about the phenomenology of consciousness. The approach to be discussed is that of *auto-meta-management* - where the first-order process meta-manages itself; illustrated in Figure 2.

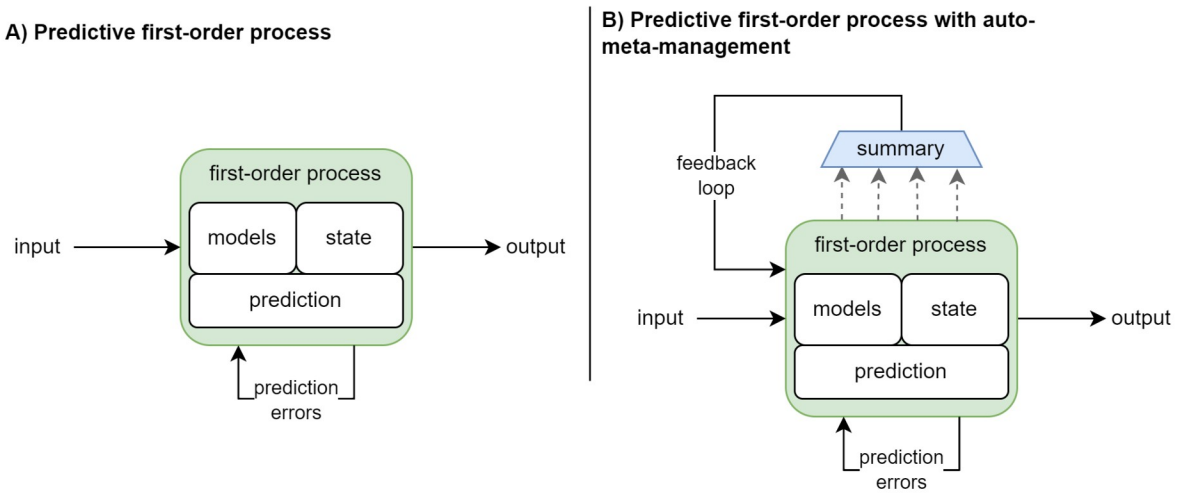

Figure 2 - First-order predictive behavioural control processes. *(A) A standard predictive process uses sensory input to predict causal latent states, and to predict suitable behaviours in response. Errors between expected and observed input and output are used to refine predictions, updating internal state. (B) With the addition of a meta-management feedback loop, the same predictive process can use predictions about its own internal state to refine its computational behaviour during deliberation. The feedback loop works by capturing a dimensionally reduced summary of the process's internal state as it operates, and makes it available as an additional input that can be modelled like any other input.*

There is increasing evidence that the sensorimotor system of the primate brain operates in a predictive fashion (Friston, 2010; Clark, 2013 and 2019; Kilner, Friston, Frith, 2007). Sensory interpretation is performed by inference of the latent states that would have generated the sensory signals, mediated by prediction errors between expected versus actual sensory signals. Motor production is produced in a similar way. The brain predicts the best outcome that would meet its

immediate needs. The sensorimotor system then predicts the best set of motor commands that would achieve that outcome. Errors between expected outcome and actual outcome are used to adjust those motor commands. Both processes operate by learning a model of the cause-effect relationships present in both the external environment and within the body of the organism. Through the iterative process of inference, prediction error, and inference adjustment, the model can be used in both familiar and novel circumstances to predict the most likely interpretation of the latent cause of sensory signals, and the behaviour most likely to meet the needs of the organism. In effect, and ignoring all the other brain processes for a moment, an organism's intelligent behaviour can be seen as an emergent property of a mechanism for learning cause-effect models and a mechanism for prediction based on those models.

In a brain lineage that has already evolved or has the potential to evolve the ability to distinguish between objects, to model the cause-effect relationships of those objects, and to infer appropriate actions while taking into consideration those models, there is only one additional evolutionary step required to enable auto-meta-management. That step is for the observation of the first-order cognitive state discussed above to be made available as a first-order sense. As illustrated in Figure 2(B), this can be achieved through a *meta-management feedback loop*, where a high-level dimensionally reduced summary is produced from a sampling of selected brain state and made available as a first-class sense to the same first-order processing systems that receive other endogenous and exogenous senses.

Effective meta-management of deliberative processing emerges from that architecture, in exactly the same way that effective behaviour w.r.t. to the body needs and environmental situations emerges from having sufficient senses and learning processes in relation to those. By having access to its own internal state, the first-order process is able to modulate its own

behaviour. By constructing causal models of its own internal processing behaviours, their relation to subsequent internal processing states, and to ultimate outcomes, it can use that as additional contextual information in producing its behaviours at any given moment. For example, to avoid processing behaviours that have historically led to unwanted outcomes.

Meta-management solves the problem of brain state becoming dissociated from immediate physical needs by directly modelling how that brain state relates to physical needs. However, it also enables such a dissociation to persist where appropriate. Cognitive state space can become a world unto itself, with its own complex structure and goals. In other words, it enables *abstract thought*.

Meta-management as Consciousness

I claim that the meta-management feedback loop and the emergence of auto-meta-management are not just the underlying mechanisms of deliberative thought, but that those same mechanisms underlay conscious awareness. More specifically, I claim that they explain the *contents of consciousness*, a topic that I shall now focus upon.

The meta-management theory of consciousness explains a number of key phenomenological aspects of conscious experience. I discuss three such key aspects here:

Limited. Conscious contents is limited to only certain classes of information within the brain state. It is now well known that many cognitive processes occur without conscious involvement (Oakley & Halligan, 2017; Linser & Goschke, 2007; Wegner, 2003 & 2002). For example, even when we have conscious awareness it is often only of the final product of inference, rather than of how that inference occurred (Oakley & Halligan, 2017). This is a natural result of avoiding an infinite regress on the size of the brain as discussed earlier, and as a result of evolution favouring more efficient solutions - only the minimal amount of information needed

for effective meta-management should be made available through the meta-management feedback loop.

Looks Through. A common observation of conscious awareness is that conscious contents "look through" to first-order states such as those of sensory perceptions, goals, and thought (Siewert, 2004, 2022; Siegel, 2021; Crane, 2009). In contrast, first-order states represent themselves: a sensory perception state captures the contents of a sensory perception; a goal state captures the contents of a goal; a thought state captures the content of a thought. Logically, you might expect that a conscious state should capture some sort of "conscious content" that is distinctly different from other sensory perceptions, goals, and thoughts; so why is that not the case? There are two comments in answer to this question. Firstly, the "looks through" nature of conscious content is explained by the fact that the purpose of the meta-management feedback loop is to capture the state of the "main goings on" within the brain. The main goings on at any given moment in time are usually in relation to sensory perceptions, goals, and thoughts. Thus the feedback loop captures a higher-order representation of those very same sensory perceptions, goals, and thoughts - it "looks through" to those states. Secondly, as a corollary, conscious content does not in fact exactly look through to the original first-order state. Rather, it offers a) the *result* of processing in relation to that first-order state, that simultaneously is b) reduced in scope and precision, and c) is enriched with additional information not available from the origin of the first-order state. An example of point (c) is that it is virtually impossible for a neurotypical individual to observe the face of a loved one without recognizing them as their loved one.

Timely. There is an apparent paradox in the timing of conscious awareness. In general, we have conscious awareness of events as they unfold in real time, and we appear to use that feature of conscious awareness in order to consciously respond in real time. However, detailed

experiments have found a lag from the moment of neurological evidence of a decision having been made to the individual being consciously aware of the decision (Libet, 1985 & 2004; Haynes, 2013; Soon et al 2008). If conscious awareness occurs *after the fact*, as it were, it begs the question of the purpose of consciousness. This has even been used as an argument for *epiphenomenalism*, the theory that consciousness has no causal effect on our actions. The meta-management theory of consciousness provides a simple answer. Conscious content is indeed constructed *after the fact* - the meta-management feedback loop only provides its contents as a result of the first-order processing that it monitors. This is acceptable from the point of view of meta-management, which has the goal of modulating the overall trajectory of brain state, without needing to be involved with the fine details. But conscious awareness is also very much causal, just in a specific way: conscious content is used *in response to* the fine-grained goings on within the brain.

Relation to Existing Theories

The meta-management theory of consciousness is compatible with a number of existing theories of consciousness. Furthermore, it adds important context to those theories by providing an overarching explanation for a) the importance of conscious awareness in brain processes, and b) how those brain processes result in conscious states.

Global Workspace Theory (GWT) (Baars, 1988 & 2021; Baars and Franklin, 2007), together with Global Neuronal Workspace Theory (GNWT) (Dehaene, Sergent, and Changeux, 2003), provides a high level framework for computational processing across multiple domains in the face of ambiguity by explaining how to coordinate many sub-processes which are each focused on certain sub-problems and which may at times compete with opposing information. GWT explains that the sub-processes can be coordinated through a choke-point of a single

"blackboard" where those sub-processes compete to broadcast their outputs to and which is subsequently used as input to all other sub-processes. GNWT extends that with a specific neuronal-level mechanism for groups of neurons to "win" in that competition. Information that is broadcast to the blackboard is claimed to form the contents of consciousness, however the theories fail to explain *how* that broadcast information takes the step from being just data to conscious data. The meta-management theory of consciousness adds how the broadcast information is meta-observed, made available as a first-order sense, identified as originating from brain processes, and thus forms the content of conscious awareness.

The various variants of the Higher Order Thought (HOT) theory of consciousness have for a long time suggested the existence of higher-order representations that capture a first-order state in conjunction with an indication that the given first-order order is being consciously experienced (Rosenthal, 1997 & 2004; Carruthers, 1996, 2000 & 2005). This is an intuitively appealing theory as it reflects our experience: when we are consciously aware of some perception, goal, or thought, we are able to be consciously aware of the fact that we are consciously aware of it. If there were a well defined definition of consciousness, something to that effect would likely be part of that definition. But why? Why should the brain expend precious energy in the production of HOTs? The meta-management theory of consciousness offers that important explanation: in order that the first-order process can be monitored and controlled. HOTs themselves are the sensory signal from the meta-management feedback loop after it has been attended to and processed in relation to causal-modals that identify the origin of sensory signals. That result may itself be re-observed as part of the ongoing meta-management process, leading to the individual concluding that not only are they aware of a given sensory perception, they are aware of being aware.

Closing Remarks

What I have described so far amounts to a thought exercise. I have proposed a series of logical steps that make claims about evolutionary needs and responses, and I have made claims about the specific structural nature of the computational brain mechanisms that resulted. This can only be proven with empirical evidence. I suspect that the structural nature underlying the meta-management feedback loop and auto-meta-management should have clear neural correlates in the form of mutual information between the brain regions involved with raw sensory perceptions, the meta-management feedback loop, and the consequent first-order processing of the signals from that feedback loop.

In the absence of the required empirical evidence, I have a number of reasons to claim that the theory as presented here is a strong contender for an explanation of consciousness. Firstly, it is consistent with existing theories that already have strong support, and which have a history of identifying potential correlated brain regions and observed neurological processes. Secondly, it offers a simple and elegant explanation to many observed phenomena, not just in terms of *correlating* with those phenomena, or even just in terms of *how* those phenomena occur, but also in terms of *why* the underlying mechanisms that produce those phenomena are evolutionarily necessary in the first place. Thirdly, the theory is testable. It makes predictions about brain structure, about the relationship between different brain states, and about the timeliness and causal relations of those brain states. Lastly, it is practical. It is easily implementable in artificial computational systems, and may prove to be useful in the future advancement of artificial intelligence.

To the last point, should this theory prove even partially correct, it may have significant influence on our understanding of human intelligence. It offers a mechanism that can sustain

meaningful and productive abstract thought. It also suggests a strong link between consciousness and intelligence, and suggests that different forms of intelligence will differ in their form of consciousness, or lack thereof.

In the interests of maintaining focus and brevity, I have focused on the mechanistic aspects of consciousness, with some reference to some phenomenological aspects. This leaves an open question about the broader metaphysical question of consciousness - why the contents of consciousness as described should carry with it the "raw feels" that it does (Nagel, 1974). I believe that the meta-management theory of consciousness can be extended to cover raw feels, and in fact to all of phenomenological consciousness. However, I fear that I lack the requisite skills to fully explicate such a debate, and I hope that skilled philosophers will take up the mantle of that argument.

Conflict of Interest

The author has no competing interests to declare that are relevant to the content of this article.

Data Availability

No new data were generated or analysed in support of this research.

References

Armstrong, S., Leike, J., Orseau, L., & Legg, S. (2020). Pitfalls of Learning a Reward Function Online. *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI-20)*. https://doi.org/10.24963/ijcai.2020/221

Baars, B. J. (1988). *A cognitive theory of consciousness.* Cambridge University Press.

Baars, B. J. (2021). *On Consciousness: Science & Subjectivity - Updated Works on Global Workspace Theory.* Nautilus Press.

Baars, B. J., & Franklin, S. (2007). An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. *Neural Networks*, 20(9), 955–961. https://psycnet.apa.org/doi/10.1016/j.neunet.2007.09.013

Beaudoin, L. (1994). Goal processing in autonomous agents [PhD thesis, The University of Birmingham]. https://citeseerx.ist.psu.edu/document?doi=382135c4379c08253810ef8f5823c469af6b69df

Benjamin, A.S., Bjork, R.A., & Schwartz, B.L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, 127(1), 55-68. https://psycnet.apa.org/doi/10.1037/0096-3445.127.1.55

Carruthers, P. (1996). *Language, Thought and Consciousness.* Cambridge University Press.

Carruthers, P. (2000). *Phenomenal Consciousness: a naturalistic theory.* Cambridge University Press.

Carruthers, P. (2005). *Consciousness: essays from a higher-order perspective.* Oxford University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–204. https://doi.org/10.1017/s0140525x12000477

Clark, A. (2019). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, 9, 1–15. https://doi.org/10.1080/00048402.2019.1602661

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind and Language*, 8, 487–519. https://doi.org/10.1111/j.1468-0017.1993.tb00299.x

Cleeremans, A. (2007). Consciousness: the radical plasticity thesis. In R. Banerjee, B. K. Chakrabarti (Eds.), *Progress in Brain Research*, 168, 19-33. Elsevier. https://doi.org/10.1016/S0079-6123(07)68003-0

Cleeremans, A., Timmermans, B., Pasquali, A. (2007). Consciousness and metarepresentation: A computational sketch. *Neural Networks*, 20(9), 1032-1039, https://doi.org/10.1016/j.neunet.2007.09.011

Cleeremans, A. (2019). The mind is deep. In A. Cleeremans, V. Allakhverdov, & M. Kuvaldina (Eds.), *Implicit learning: 50 years on* (pp. 38–70). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315628905-3

Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J-R., Muñoz-Moldes, S., Vuillaume, L., & de Heering, A. (2020). Learning to Be Conscious. *Trends in Cognitive Sciences*, 24(2), 112-123, https://doi.org/10.1016/j.tics.2019.11.011

Crane, T. (2009). Intentionalism. In B. McLaughlin, & A. Beckermann (Eds.), *The Oxford Handbook to the Philosophy of Mind* (pp. 474–93). Oxford University Press. (Full text: https://philpapers.org/rec/CRAI-17)

Dehaene, S., Sergent, C. & Changeux, J. P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *The Proceedings of the National Academy of Sciences (PNAS)*, 100, 8520-5. https://doi.org/10.1073/pnas.1332574100

Fernandez Cruz, A.L., Arango-Muoz, S., Volz, K.G. (2016). Oops, scratch that! Monitoring one's own errors during mental calculation. *Cognition*, 146, 110-120. https://doi.org/10.1016/j.cognition.2015.09.005

Fleming, S.M., Dolan, R.J., & Frith, C.D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11, 127–138. https://doi.org/10.1038/nrn2787

Gleitman, H., Fridlund, A. J., & Reisberg, D. (2004). *Psychology* (6th ed.). Norton.

Godfrey-Smith, P. (2016). *Other Minds: The octopus and the evolution of intelligent life.* Farrar, Straus and Giroux.

Haynes, J-D. (2013). Beyond Libet: Long-Term Prediction of Free Choices from Neuroimaging Signals. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the Will* (pp. 60-72). Oxford University Press.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science.* MIT Press

Karmiloff-Smith, A. (1994). Précis of Beyond modularity: A developmental perspective on cognitive science. *Behavioral and Brain Sciences*, 17(4), 693-707. https://doi.org/10.1017/S0140525X00036621

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, 8(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2

Lau H. (2007). A higher order Bayesian decision theory of consciousness. *Progress in Brain Research*, 168, 35–48. http://doi.org/10.1016/S0079-6123(07)68004-2

Lazaridis, A., Fachantidis, A., & Vlahavas, I. (2020). Deep Reinforcement Learning: A State-of-the-Art Walkthrough. *Journal of Artificial Intelligence Research*, 69, 1421-1471. https://doi.org/10.1613/jair.1.12412

Libet, B. (1985). Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *Behavioral and Brain Sciences*, 8(4), 529–566. https://psycnet.apa.org/doi/10.1017/S0140525X00044903

Libet, B. (2004). *Mind Time: The Temporal Factor in Consciousness.* Harvard University Press.

Marković, D., Goschke, T., & Kiebel, S.J. (2021). Meta-control of the exploration-exploitation dilemma emerges from probabilistic inference over a hierarchy of time scales. *Cognitive, Affective & Behavioral Neuroscience*, 21(3), 509–533. https://doi.org/10.3758/s13415-020-00837-x

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review,* 83(4), 435–450. https://doi.org/10.2307/2183914

Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition*, 117(2), 182–190. https://doi.org/10.1016/j.cognition.2010.08.010

Paulin, M.G., & Cahill-Lane, J. (2021), Events in Early Nervous System Evolution. *Topics in Cognitive Science*, 13, 25-44. https://doi.org/10.1111/tops.12461

Rosenthal, D. M. (1997). A Theory of Consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness: Philosophical Debates* (pp. 729-753). *MIT Press/Bradford Books.* (https://davidrosenthal.org/DR-A-Theory.pdf)

Rosenthal, D. M. (2004). Varieties of higher-order theory. In R. Gennaro (Ed.), *Higher-Order Theories of Consciousness: An Anthology* (pp. 17-44). John Benjamins. http://doi.org/10.1075/aicr.56.04ros

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003.

Schroeder, T. (2020). Desire. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition). https://plato.stanford.edu/archives/sum2020/entries/desire/

Siegel, S. (2021). The Contents of Perception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). https://plato.stanford.edu/archives/fall2021/entries/perception-contents/

Siewert, C. (2004). Is Experience Transparent? *Philosophical Studies*, 117(1/2), 15–41. https://doi.org/10.1023%2FB%3APHIL.0000014523.89489.59

Siewert, C. (2022). Consciousness and Intentionality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition). https://plato.stanford.edu/archives/sum2022/entries/consciousness-intentionality/

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: consciousness as a unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 1412–1423. https://doi.org/10.1098%2Frstb.2011.0421

Sloman, A. (1998). Damasio, Descartes, alarms and meta-management. *IEEE International Conference on Systems, Man, and Cybernetics*, 3, 2652-2657. https://doi.org/10.1109/ICSMC.1998.725060

Sloman, A. (2008). Varieties of Metacognition in Natural and Artificial Systems. In M. T. Cox & A. Raja (Eds.), *Metareasoning: Thinking about Thinking.* MIT Press. https://doi.org/10.7551/mitpress/9780262014809.003.0020

Smithies, D., & Weiss, J. (2019). Affective Experience, Desire, and Reasons for Action. *Analytic Philosophy*, 60(1), 27-54. https://doi.org/10.1111/phib.12144

Soon, C., Brass, M., Heinze, H-J., & Haynes, J-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11, 543–545. https://doi.org/10.1038/nn.2112

Spoerer, C. J., Kietzmann, T. C., Mehrer J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, 16(10), e1008215. https://doi.org/10.1371/journal.pcbi.1008215

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed). MIT press.

van Bergen, R. S., & Kriegeskorte, N. (2020). Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65(176-193). https://doi.org/10.1016/j.conb.2020.11.009

Winkielman, P., & Schooler, J. (2012). Consciousness, metacognition, and the unconscious. In S.T. Fiske, & C.N. Macrae (Eds.), *The SAGE Handbook of Social Cognition.* SAGE Publications Ltd. https://doi.org/10.4135/9781446247631.n4