Grounded Metacognitive Architectures For Machine Consciousness

Ron Chrisley[0000−0002−1177−1789]

Centre for Cognitive Science, Sackler Centre for Consciousness Science, and
Department of Informatics
University of Sussex, Falmer BN1 9QJ, United Kingdom
ronc@sussex.ac.uk; http://www.sussex.ac.uk/profiles/476

Abstract. Multiple approaches to machine consciousness emphasise the importance of metacognitive states and processes. A considerable number of cognitive systems researchers prefer architectures that are not classically symbolic, and in which learning, rather a priori structure, is central. But it is unclear how these grounded architectures can support metacognition of the required sort. To investigate this possibility, a basic design sketch of such an architecture is presented.

Keywords: Metacognition · Sub-symbolic Computation · Cognitive Architecture · Symbol Grounding · Neural Networks · Machine Consciousness.

1 Motivation

1.1 Metacognition and Machine Consciousness

Cognition is used here, rather idiosyncratically, to mean processes and states that enable an agent to act appropriately with respect to a subject matter (especially one that is distal and/or abstract). *Metacognition*, here, is cognition about cognition: processes and states that enable a cognitive agent to act appropriately with respect to (especially, but not exclusively, its own) cognitive processes and states. A paradigmatic example would be beliefs about (and subsequent reasoning about, and action on) beliefs. There are at least three distinct reasons why modelling metacognition might be crucial to modelling consciousness:

1. Metacognition might be central to consciousness itself. For example, most if not all higher-order thought theories of consciousness (e.g. [8]) explicitly stipulate that what it is for a mental state to be a conscious mental state is that it is represented by a higher-order (viz, metacognitive) state.
2. It might be that the proper explanation of some purported properties of consciousness proceeds by way of explaining why conscious agents are disposed to attribute those properties to their own conscious states. Such explanations are thus inherently metacognitive.
3. It might play a role in explaining *cognitive phenomenology*, as opposed to sensory phenomenology.

(More is said about 2 and 3 in section 4.) If any of these lines of reasoning are correct, metacognitive states and processes are of central importance to computational models of human-like consciousness (reasons 2 and 3) or any consciousness at all (reason 1).

Observation 1: Artificial intelligence approaches to modelling metacognition to date have almost exclusively employed classical symbolic architectures. It has usually been presupposed that not only the second-order states, but even the lower-level states that are the objects of metacognition are classically symbolically structured.

1.2 Grounded Architectures

On the other hand, a considerable number of cognitive systems researchers eschew such architectures in favour of ones that are *grounded*, in that they explain, rather than presuppose, behaviour and capacities typically believed to require symbolic representation and processing. Where such architectures invoke representation and processing at all, it is sub-symbolic, with a concrete semantics that concerns (at least in the first instance) the dynamic here-and-now of sensorimotor engagement, rather than context-free, abstract, fully conceptual ways of representing the world, such as those (supposedly) employed by the theorist. Such architectures typically invoke learning, for (hopefully familiar) reasons that cannot be reiterated here.

Observation 2: Those employing grounded architectures have tended to focus, understandably, on modelling "lower-level" cognitive phenomena such as perception and action selection, and have tended to shy away from modelling higher cognitive capacities, of which metacognition is a paradigmatic example.

1.3 Toward Grounded Metacognitive Architectures

There is a *prima facie* tension between Observations 1 and 2: Metacognitive modelling has tended not to be grounded; grounded architectures are rarely applied to model metacognitive phenomena. Thus there is a relative lack of of architectures for metacognition that are not classically symbolic, and in which learning, rather a priori structure, is central.

However, metacognition has great potential for improving the abilities of grounded architectures: not just for improving learning, but also for playing a role in providing (otherwise absent) sophisticated representational features, such as systematicity, productivity, logical quantification, modality (possibility/necessity), etc. Although there is a growing, relatively recent literature on the related topic of *metalearning* (e.g., [9, 7]), the concern there has more to do with optimising machine learning, and there is little or no discussion of topic at hand: machine consciousness.

To address this deficiency, a design sketch of such an architecture is presented in what follows Even in such an undeveloped state, the sketch can be used to clarify the kinds of architectures in question, the design challenges they face, and the requirements they must meet to assist in an explanation of some of the most difficult features of consciousness.

It should be emphasised that the sketch is incomplete not only as concerns detail (depth), but also as concerns the number of components included (breadth). Rather than a full cognitive architecture, what is presented here is only what is necessary to illustrate the proposed metacognitive design.

2 A Design for Grounded Metacognition

2.1 The First-Order Subsystem

Although the metacognitive subsystem described below can be applied to a very wide range of first-order systems, including (passive) classification/pattern recognition networks, it will be easier to identify its distinctive features in the context of an action-involving first-order system.

Consider a robot in an effectively 2D environment of simple coloured polygons. The robot is static except that it can move its single camera to fixate on points in the 2D field. The result R of fixating on point (x, y) is that the sensors take on a particular value s out of a range of possible values S. That is, $R(x, y) = s \in S$.

The visual environment is dynamic in that at any given time one or more polygons may disappear or appear. However, the time between such changes is long relative to the learning described below: learning in response to a change will be assumed to have converged before the next change takes place. (The risks of this assumption can be mitigated by giving the robot a change detection component, that indicates when such an alteration of the 2D field has taken place, and which may be used to re-initialise the learning process, if necessary).

The robot learns a forward model F with parameters (weights) $\bar{w}$ from points of fixation (x, y) to expected sensory input s at the fixated location: $F_{\bar{w}}(x, y) = s \in S$.

2.2 The Metacognitive Subsystem

Since F, in effect, manifests the robot's "beliefs" about the state of its environment, and F is parameterised by $\bar{w}$, representations of ("beliefs" about) $\bar{w}$, and operations/transformations on $\bar{w}$, amount to a kind of metacognition.

The metacognitive system, then, can also be realised in a forward model, M. But the simplicity of the first-order forward model is inadequate for metacognition, for two reasons.

First, it is not feasible to represent metacognitive actions directly in terms of their effects, as was done with $R(x, y)$; rather they should be conceived as functions from an input weight state, and an operator on that state, to an output weight state. To illustrate, a similar representation for the first-order model (but not one being used in the architecture described in this paper) would be a function R from the current point of fixation (x, y) and an operator O to a new point of fixation (x', y'); for example: $R(x, y, O_{u,v}) = (x + u, y + v)$. Thus, for the metacognitive subsystem, we have: $M_{\bar{j}}(\bar{w}, O_k) = \bar{w'}$.

Of course, the operators O_k in the previous equation have not yet been defined, which is the result of a second difference between the first-order and metacognitive subsystems: while we have intuitive ideas as to what should constitute the primitive action basis set for many first-order cognitive systems, it is less clear what such a set should be for metacognition. This problem is confounded by the fact that metacognitive operators, by their very nature, are likely to be highly dependent on the implementation details of the first-order subsystem, and thus relatively subjective. A key part of a grounded metacognitive architecture, then, will be some principled scheme for *deriving* or *learning* such a basis set of operators, or actions. To this end, three strategies for the acquisition of metacognitive operators are under consideration (that individuate such operators by reference to weight clusters, sensory clusters, and inflections in error history, respectively), but there is no space here to detail these proposals. But a further point can be made: once such strategies are under consideration, a case can also be made for learning, rather than a priori imposing, the basis set of actions for the *first-order* subsystem. It may be that whatever scheme for learning is arrived at for learning metacognitive operators might be a good starting point for acquiring the first-order basis actions.

Despite the differences, there is a parallel between the first-order and metacognitive subsystems in that in both, a forward model of the impact that operators or actions have on their domains (sensations or weights, respectively) can be used, for example, to assist in action selection [2]. Details of how this might be done in the case of metacognition are given below.

Construed this way, metacognitive networks will be of much higher dimensionality than first-order ones (since one of the dimensions of the metacognitive network will be equal to the number of parameters in the first-order network). Perhaps this fact explains why such approaches to metacognition have not seen much or any development until now, when hardware advances make at least modest versions of this architecture tractable. On the other hand, an important part of future would will be to justify the considerable computational cost that this will incur: metacognition should earn its keep.

3 Learning to Use Metacognitive Actions

Once a set of metacognitive operators are in place (by applying, say, one of the three strategies mentioned above), an agent can learn a forward model $M_{\bar{j}}(\bar{w}, O) = \bar{w'}$ over first-order weight states. That is, the agent can learn the expected effects that performing one of these operators (that is, taking metacognitive action) on a given weight set would have, just as an agent can learn what effects its first-order actions have on its sensory input. The key difference is that the key contingencies to be learned are not (just) “within-level”, as with learning the relation between movements and sensory input. Rather, cross-level contingencies can also be learned. Since forward models can, in some situations, be inverted to yield inverse models, a space of several kinds of cross-level models can generated, and learned:

- 2nd order forward/1st order forward models: If I were to make this change to my belief state, then what would I expect to see if I were to take this action? That is, given a metacognitive operator O, and an first-order action of interest a, calculate $s = F_{\bar{w}'}(a)$, where $\bar{w'} = M_{\bar{j}}(\bar{w}, O))$.
- 2nd order forward/1st order inverse model: If I were to make this change to my belief state, what action would then be expected to yield a given sensory input? That is, given a metacognitive operator O, and a first-order sensory input of interest s, calculate $a = F^{-1}_{\bar{w}'}(s)$, where $\bar{w'} = M_{\bar{j}}(\bar{w}, O))$.
- 2nd order inverse/first-order inverse models: What metacognitive operator/operator sequence, if any, can get me to a weight state such that there is an action that can yield a sensory input of interest? (Asking this question might be particularly useful in contexts that can benefit from creative problem solving.) That is, given a first-order sensory input of interest s, calculate $O = M^{-1}_{\bar{j}}(\bar{w}, \bar{w'})$, such that for some a, $s = F_{\bar{w'}}(a))$.

This learning can be of either the machine learning/neural network variety, or more classically symbolic, even though what is being learned about may be, as it is here, neural network states. But there may be some explanatory benefits (e.g. concerning origin, or a general parsimony) in having M realised in the same kind of neural network, say, as F.

An important possibility that can only be noted in passing here is that metacognitive learning may be facilitated by explicitly preferring first-order models that support metacognitive operators (via adding an error term to $F's$ objective function, say) . Such dovetailing may also occur indirectly (via something akin to the Baldwin Effect).

4 The Upshot for Machine Consciousness

4.1 A Revisionist Solution to the Hard Problem

Earlier work ([10, 5, 6]) has argued for an indirect way for machine consciousness to tackle the Hard problem of consciosuness [1], conceived here as the problem of explaining how can something physical can have the properties (intrinsicness, ineffability, etc.) we typically ascribe to our conscious experiences. The first step is to account for the aspects of consciousness that proponents of the Hard problem concede are not Hard (attention, cognition, memory, learning, etc). The second step is to show why it is that systems with complex combinations of those capacities tend to believe certain things about the states that they are in: e.g., that their states have intrinsic, qualitative character, of the kind for which a Hard problem would arise. This two-step approach to machine consciousness requires metacognitive models, since the belief tendencies in the second step are second-order. One cannot provide this explanation just by stipulating that a system has the second-order beliefs in question: one must explain why a given state constitutes such a belief, and why the system has a tendency to hold (and resist revising) such beliefs. Thus, such explanations require a grounded metacognitive model. It is proposed that the kinds of architectures (involving F, O_k, M, etc.), in that are models of systems that have beliefs about their own beliefs, are a first step toward explaining the particular kinds of meta-belief that play a role in defining the Hardness of consciousness.

4.2 Cognitive Phenomenology

In addition to this role, a further relation between these architectures and machine consciousness concerns a possible explanation of cognitive phenomenology (the experience of being in a given cognitive state) as opposed to sensory (or sensorimotor phenomenology). There is no consensus that there is even such a thing as cognitive phenomenology, but prior work on how forward models like F can characterise the content of sensory experience ([3, 4]), together with the parallels between F and M, suggest a simple account. Just as the content of sensory phenomenology at a time may be given by the set of sensorimotor expectations F realises at that time, so also the content of cognitive phenomenology at a time may be given by the set of metacognitive expectations M realises at that time. Just as the content of sensorimotor experience may consist in the set of answers to questions such as "What input would I receive were I to move my eyes this way?", the content of "cognitive experience" may consist in the set of answers to questions such as "What cognitive state would I be in were I to apply this metacognitive operator?".

The plausibility of this account of cognitive phenomenology increases when it is made clear that this kind of expectation-based account of experiential content is not restricted to expectations manifested in the lowest-level forward model (of either the first-order (F) or metacognitive (M) subsystems). Each of these subsystems can contain an abstraction hierarchy, consisting of more abstract (as opposed to higher-order, or meta-) representations of their inputs and outputs, and the relations between them.

For example, once $F_{\bar{w}}(a^0) = s^0$ is learned for lowest (0th) level action and sensory schemes, action and sensory hierarchies $(A_{i+1}(a^i) = a_{i+1}, S_{i+1}(s^i) = s_{i+1})$ can themselves be learned, grounded by F. These can be thought of as traditional pattern classifier networks, in which higher levels produce more abstract representations of the lowest level inputs (a^0 and s^0, respectively). These hierarchies generate a corresponding hierarchy of (all first-order, but of increasing abstraction) forward models F^i (the superscript here being logically distinct from the -1 superscript used before to denote an inverse model). This allows for a characterisation of experiential content of varying abstraction, so that experience need not only be in terms of basic motor commands and sensory inputs, but also actions such as *open-the-door* and inputs such as *book-shaped-object*.

The above first-order example illustrates, by extension, comparable possibilities for the metacognitive system. By learning abstraction hierarchies for weight states and metacognitive operators, it becomes possible to model cognitive phenomenology that goes beyond the most concrete level to more abstract characterisations, involving experience of, say, *changing one's mind* to a *more cautious appraisal of the situation*, etc.

The fact that these abstraction hierarchies can cross-cut the first-order/meta hierarchy implies a wide range of architectures of increasing complexity. One dimension of variation concerns how the first-order hierarchy of models F^i is implemented. One way is for each of the F^i to be implemented with independent sets of weights, resulting in a corresponding set of distinct M^i to operate over them. Besides requiring a large amount of resources, ensuring integrity between changes on all these levels looks to be a daunting (and expensive) task. Another approach would be to implement only F^0, with $F^{i>0}(a^i) = s^i$ approximated by $S(F^0(A^{-i}(a^i)))$. That is, the action hierarchy is inverted to yield a 0th-order representation of the abstract action a^i; this low-level action is fed into the 0th-order forward model to yield a low-level expectation of sensory input, which is in turn fed into the sensory pattern hierarchy to yield an abstract representation of that low-level sensory input, to yield the required abstract expected input, s^i. An analogous dimension of variation exists for the hierarchy of metacognitive models M^i: the expensive option of having distinct implementations, or a virtual option of grounding the M^i in M^0 in a manner similar to what was just described for the first-order subsystem.

References

1. Chalmers, D.: The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press, Oxford (1996)
2. Chrisley, R.: Cognitive map construction and use: A parallel distributed processing approach. In: Touretzky, D., Elman, J., Sejnowski, T., Hinton, G. (eds.) Connectionist Models: The Proceedings of the 1990 Connectionist Models Summer School. Morgan Kaufmann, San Mateo (1990)
3. Chrisley, R., Parthemore, J.: Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience. Journal of Consciousness Studies **14**, 44–58 (2007)
4. Chrisley, R.: Synthetic phenomenology. International Journal of Machine Consciousness **01**(01), 53–70 (2009)
5. Chrisley, R., Sloman, A.: Functionalism, revisionism, and qualia. APA Newsletter on Philosophy and Computers **16**(1), 2–13 (December 2016)
6. Chrisley, R., Sloman, A.: Architectural requirements for consciousness. In: Chrisley, R., Müller, V.C., Sandamirskaya, Y., Vincze, M. (eds.) EUCognition 2016: Cognitive Robot Architectures. vol. 1855, pp. 31–36. CEUR Workshop Proceedings (June 2017)
7. Lemke, C., Budka, M., Gabrys, B.: Metalearning: a survey of trends and technologies. Artificial Intelligence Review **44**(1), 117–130 (Jun 2015)
8. Rosenthal, D.: Two concepts of consciousness. Philosophical Studies **49**, 329–359 (01 1986). https://doi.org/10.1007/BF00355521
9. Schaul, T., Schmidhuber, J.: Metalearning. Scholarpedia **5**(6), 4650 (2010). https://doi.org/10.4249/scholarpedia.4650, revision #91489
10. Sloman, A., Chrisley, R.: Virtual machines and consciousness. Journal of Consciousness Studies **10**(4-5), 133–172 (January 2003)