Architectural Requirements for Consciousness

Ron Chrisley
Centre for Cognitive Science (COGS),
Sackler Centre for Consciousness Science,
and Department of Informatics
University of Sussex
Brighton, United Kingdom
Email: ronc@sussex.ac.uk

Aaron Sloman
Department of Computer Science
University of Birmingham
Birmingham, United Kingdom
Email: axs@cs.bham.ac.uk

***Abstract*—This paper develops, in sections I-III, the virtual machine architecture approach to explaining certain features of consciousness first proposed in [1] and elaborated in [2], in which particular qualitative aspects of experiences (qualia) are proposed to be particular kinds of properties of components of virtual machine states of a cognitive architecture. Specifically, they are those properties of components of virtual machine states of an agent that make that agent prone to believe the kinds of things that are typically believed to be true of qualia (e.g., that they are ineffable, immediate, intrinsic, and private). Section IV aims to make it intelligible how the requirements identified in sections II and III could be realised in a grounded, sensorimotor, cognitive robotic architecture.**

I. Introduction

Those who resist the idea of a computational, functional, or architectural explanation of consciousness will most likely concede that many aspects surrounding consciousness *are* so explicable (the so-called "easy problems" of consciousness [3]), but maintain that there are core aspects of consciousness having to do with phenomenality, subjectivity, etc. for which it is Hard to see how a computational explanation could proceed. A typical way of characterising this "Hard core" of consciousness employs the concept of *qualia*: "the introspectively accessible, phenomenal aspects of our mental lives" [4]. Surely there can be no computational explanation of qualia?

This paper develops the virtual machine architecture approach to explaining certain features of consciousness first proposed in [1] and elaborated in [2], in which qualia, understood as particular qualitative aspects of experiences, are proposed to be particular kinds of properties of components of virtual machine states of a cognitive architecture. Specifically, they are those properties of components of virtual machine states of agent A that make A prone to believe:

1) That A is in a state S, the aspects of which are knowable by A directly, without further evidence (immediacy);
2) That A's knowledge of these aspects is of a kind such that only A could have such knowledge of those aspects (privacy);
3) That these states have these aspects intrinsically, not by virtue of, e.g., their functional role (intrinsicness);
4) That these aspects of S cannot be completely communicated to an agent that is not A (ineffability).

Our emphasis on beliefs concerning these four properties (immediacy, privacy, intrinsicness and ineffability), follows the analysis in [5] in taking these properties to be central to the concept of *quale* or *qualia*. But whereas [5] understands this centrality to imply that the properties themselves are conditions for falling under the concept, we understand their centrality only in their role of *causally determining* the reference of the concept. Roughly, qualia are not whatever has those four properties; rather, qualia are whatever is (or was) the cause of our qualia talk. And if we do know anything about the cause of our qualia talk, it is this: it makes us prone to believe that we are in states that have those four properties.

A crucial component of our explanation, which we call the Virtual Machine Functionalism (VMF) account of qualia, is that the propositions 1-4 need not be true in order for qualia to make A prone to believe those propositions. In fact, it is arguable that nothing could possibly render all of 1-4 true simultaneously [5]. But on our view, this would not imply that there are no qualia, since for qualia to exist it is only required that that agents that have them be prone to believe 1-4, which can be the case even when some or all of 1-4 are false.

It is an open empirical question whether, in some or all humans, the properties underlying the dispositions to believe 1-4 have a unified, systematic structure that would make them a single cause, and that would thereby make reference to them a useful move in providing a causal explanation of such beliefs. Is "qualia" more like "gold", for which there was a well-defined substance that was the source of mistaken, alchemical talk and beliefs about gold? Or is "qualia" more like "phlogiston", in that there is no element that can be identified as the cause of the alchemists' mistaken talk and beliefs that they expressed using the world "phlogiston"? These are empirical questions; thus, according to the VMF account of qualia, it is an open empirical question whether qualia exist in any particular human. By the same token, however, it is an open *engineering* question whether, independently of the human case, it is possible or feasible to design an artificial system that a) is also prone to believe 1-4 and b) is so disposed because of a unified, single cause. Thus, it is an open engineering question whether an artificial system can be constructed to have qualia. This paper goes some way toward getting clear on how one would determine the answer to that engineering question.

Section II notes the general requirements that must be in place for a system to believe 1-4, and then sketches very briefly, in section III, an abstract design in which the propensities to believe 1-4 can be traced to a unified virtual machine structure, underwriting talk of such a system having qualia.

II. General architectural requirements for having qualia

General requirements for meeting constraints 1-4 include being a system that can be said to have beliefs and propensities to believe, as well as what those properties themselves require. Further, having the propensities to believe 1-4 in particular requires the possibility of having beliefs about oneself, one's knowledge, about possibility/impossibility, and other minds. At a minimum, such constraints require a cognitive architecture with reactive, deliberative and meta-management components [1], with at least two layers of meta-cognition: (i) detection and use of various states of internal virtual machine components; and (ii) holding beliefs/theories about those components.

III. A qualia-supporting design

A little more can be said about the requirements that 1-4 might impose on a cognitive architecture.

1) A propensity to believe in immediacy (1) can be explained in part as the result of the meta-management layer of a deliberating/justifying but resource-bounded architecture needing a basis for terminating deliberation/justification in a way that doesn't itself prompt further deliberation or justification.
2) A propensity to believe in privacy (2) can be explained in part as the result of a propensity to believe in immediacy (1), along with a policy of *normally* conceiving of the beliefs of others as making evidential and justificatory impact on one's own beliefs. To permit the termination of deliberation and justification, some means must be found to discount, at some point, the relevance of others' beliefs, and privacy provides prima facie rational grounds for doing this.
3) A propensity to believe in intrinsicness (3) can also be explained in part as the result of a propensity to believe in immediacy, since states having the relevant aspects non-intrinsically (i.e., by virtue of relational or systemic facts) would be difficult to rectify with the belief that one's knowledge of these aspects does not require any (further) evidence.
4) An account of a propensity to believe in ineffability (4) requires some nuance, since unlike 1-3, 4 is in a sense true, given the causally indexical nature of some virtual machine states and their properties, as explained in [2]. However, properly appreciating the truth of 4 requires philosophical sophistication, and so its truth alone cannot explain the conceptually primitive propensity to believe it; some alternative explanations must be offered, but it is not possible to do so here.

IV. Cognitive Architecture, not Cognitivist Architecture?

Given the anti-cognitivist, anti-representational, anti-symbolic, embodied, enactivist, etc. inclinations of many in the EUCognition community, the foregoing may be hard to accept given its free use of representational and computational notions such as belief, deliberation, justification, etc. The rest of this paper, then, is an attempt at an in-principle sketch of how one can have a grounded, dynamic, embodied, enactive(ish) cognitive architecture that nevertheless supports the notions of belief, inference, meta-belief, etc. that this paper has just maintained are necessary for the subjective, qualia aspect of consciousness, if not all aspects of consciousness.

This motivation is not strictly (that is, philosophically) required, for two reasons:

- First, our self-appointed philosophical opponents do not claim that the "easy problems" of consciousness cannot be solved physicalistically, or even computationally. Thus, in giving our explanation of the "Hard core" of consciousness, qualia, we can help ourselves to any of the capacities that are considered to fall under the "easy problems", which is the case for all of the requirements we identified in sections II and III.
- Second, an aspect a of a cognitive architecture A can be *of the same kind* as an aspect b of a distinct cognitive architecture B, even if B is capable of the sorts of beliefs mentioned in 1-4 because of possessing b, and A, despite having a, is not capable of having those sorts of beliefs. On our account, A might still have qualia by virtue of having a; this is why our account does not, despite appearances, over-intellectualize qualia, and is instead consistent with, e.g., the empirical possibility that animals and infants have qualia.

However, showing how architectures that *do* have the kinds of beliefs mentioned in 1-4 *can* be constructed out of grounded sensorimotor components is required if we are to achieve any understanding of what a system that is *incapable* of having those beliefs would have to be like for it to nevertheless warrant ascription of qualia.

This section (that is, the rest of this paper) will not have much to say about consciousness or qualia *per se*. Furthermore, the sketched architectures are likely not optimal, feasible, or even original. That there is some better way to solve the task that we use for illustrative purposes below is not to the point. The architectures and task are intended merely to act as a proof-of-concept, as a bridge between the kind of robotic systems that many in the EUCognition community are familiar or comfortable with, and the kind of robotic cognitive architecture that we have argued is required for qualia.

A. Robotic architecture, environment and task

Consider a robot that is static except that it can move its single camera to fixate on points in a 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 populated by simple coloured polygons, at most one (but perhaps none) at each fixation point (x,y). This visual environment is static during trials, although it may change from trial to trial.

The robot has learned a map M that is a discrete partition of S into a set of categories or features F (e.g., a self-organising feature map): $M(s) = f_i \in F$. In general, M is always applied to the current sensory input s, thus activating one of the feature nodes or vectors. For example, f_1 might be active in those situations in which the robot is fixating on a green circle, f_2 might be active in those situations in which the robot is fixating on a red triangle, etc.

Suppose also that the robot has the ability to detect the occurrence of a particular auditory tone. After the tone is heard, a varying visual cue (for example, a green circle) appears in some designated area of the field (the upper left corner say). The robot's task (for which it will be rewarded) is to perform some designated action (e.g. say "yes") if and only if there is something in the current visual environment (other than in the designated cue area) whose feature map classification matches that of the cue, that is: say "yes" iff $\exists(x,y) : M(R(x,y)) = M(cue)$.

There are, of course, many strategies the robot could use to perform this task. For illustrative reasons, we will consider three.

B. Strategy One: Exhaustive search of action space

The first strategy is an exhaustive search of action space. The robot performs a serial exhaustive search of the action space $R(x,y)$, stopping to say "yes" if at any point $M(R(x,y)) = M(cue)$. This requires motor activity, and is likely to take a relatively long time to perform, although it requires no "offline" preparation time. It is a "knowledge-free" solution.

C. Strategy Two: Exhaustive search of virtual action space

The second strategy is to perform an exhaustive search of a *virtual* action space.

1) Strategy Two, Version 1: Prior to hearing the tone, the robot learns a forward model E_w from points of fixation (x,y) to expected sensory input s at the fixated location: $E_w(x,y) = s \in S$. After the tone and presentation of the cue, the robot then performs a serial exhaustive search of the expectation space $E_w(x,y)$, stopping if at any point $M(E_w(i,j)) = M(cue)$. The robot then fixates on (i,j), and if $M(R(i,j)) = M(cue)$, then it says "yes". Otherwise, the search of the expectation space resumes. As this search is for the most part virtual, only occasionally requiring action (assuming E is reasonably accurate), this will be much faster than the first strategy.

2) Strategy Two, Version 2: If the idea of an exhaustive serial search of the expectation space is not considered neurally plausible enough, a a second version of the second strategy could employ a kind of content-addressable search (following ideas first presented in [6]). The difference between cue and $E(x,y)$ (or between $M(cue)$ and $M(E(x,y))$; see below) can be used as a differentiable error signal, permitting gradient descent reduction of error not in weight w space, but in visual space (which is here the same as fixation space and action space). That is (hereafter re-writing (x,y) as u), the robot can apply the Delta rule, changing u proportionally to the partial derivative of the error with respect to u:

$$\Delta u = \mu^{\frac{\partial[\frac{1}{2}(cue-E(u))^2]}{\partial u}} .$$

Since the task question is primarily about matching one of the cue categories f_i and not the cue itself, this process requires changing the robot's virtual fixation point u according to the above equation, and then checking to see if $M(E(u)) = M(cue))$. If not, u is again updated according to the Delta rule. Alternatively, one could measure the error directly in terms of differences in feature map (M) output; then the Delta rule would prescribe:

$$\Delta u = \mu^{\frac{\partial[\frac{1}{2}(M(cue)-M(E(u)))^2]}{\partial u}} .$$

In either case, this process should eventually arrive at a value u' that is a minimum in error space, although the number of iterations of changes to u required to do so will depend on a number of factors, including μ, which itself is constrained by the "spikiness" of the error space with respect to fixation points. This could result in many changes to u, but as such changes are virtual, rather than actual changes in robot fixation point, they can be performed much faster than real-time.

Standard problems with local minima apply: the fixed point in u/error space where the derivative is zero may not only not be a point for which *actual* error is zero (that is, where $M(R(u')) = M(cue)$); it may not even be a point for which *expected* error is zero (that is, where $M(E(u')) = M(cue)$). Nonetheless, u' can serve as a plausible candidate solution, which can be checked by having the robot fixate on u' via $R(u')$. If a match ($M(R(u')) = M(cue)$) is not achieved, standard neural network methods for handling local minima can be applied, if desired, to see if a better result can be obtained.

This second version of the second strategy may in some cases be more efficient than the first variation, in that it is non-exhaustive. But both verisons of the second strategy buy online performance at the price of prior "offline" exploration of the action space, and the computational costs of learning and memory.

As an aside, we note that the second version of strategy two can be used in conjunction with strategy one (or even the first verison of strategy two), in that it can suggest a heuristically-derived first guess for a real-world (or virtual) search of points in the vicinity of that guess. In the case of failure, it wouldn't be useful as it stands, it seems; since E is deterministic, when asked for a second guess after the failure of the first, strategy two would give the same recommendation again. However, it should be noted that the gradient descent method is dependent on an initial guess u, and derives candidate solutions as modifications to u. Therefore, it will give different u' answers if a different initial u is selected to seed the gradient descent process, with the new u' corresponding to the local error minimum that is closest to the new u seed chosen. Thus, search of the entire virtual (or actual) fixation point (u) space can be reduced, in theory, to a virtual search of the much smaller space of error basins in u-space. To prevent wasteful duplication of effort, there would have to be some way for the network to consider only previously-unconsidered seeds; perhaps inhibition of previously-considered seeds could achieve this.

D. Strategy Three: Learning a mapping from mappings to cues

A third strategy builds on the second strategy by employing a form of reflection or meta-cognition to guide search more efficiently. As with the second strategy, an expectational, forward model E_w is used. Note that for any given kind of cue (node or reference vector in the range of the feature map M), we can define the set P_{cue} to be all those parameter (weight) sets w for E that yield a forward model that contains at least one expectation to see that cue. That is, $P_{cue} = \forall w : \exists(x, y) : M(E_w(x, y)) = cue$.

With a network distinct from the one realising E, the robot can learn an approximation of P_{cue}. That is, the robot can learn a mapping F_{cue} from weight sets for E to {1,0}, such that $F_{cue}(w) = 1$ iff $w \in P_{cue}$. Generalising, the robot can learn a mapping F from cues and weight sets for E to {1,0}, such that $F(cue, w) = 1$ iff $w \in P_{cue}$. That is, F is a network that, given a vector w and a cue, outputs a 1 only if w parameterises a forward model E_w for which there is at least one fixation point (x, y) such that E_w "expects" cue as input after performing $R(x, y)$.

Given this, a third strategy for performing the task is to simply input the current E parameter configuration w and the cue into F, and say "yes" iff $F(w, cue) = 1$ (or, if one prefers, make the probability of saying "yes" proportional to $F(w, cue)$).

Like strategy two, strategy three spends considerable "offline", pre-task resources for substantial reductions in the time expected to complete the online task. However, unlike both strategy one and strategy two, this third strategy answers the task question *directly*: it determines whether the existential condition of the task question holds without first finding a particular fixation point that satisfies the property that the task condition (existentially) quantifies over. A drawback of this is that the robot cannot, unlike with strategy two, check its answer in the real world (except by essentially performing strategy one). But as it is essentially a lookup computation, it is very fast: no search, even virtual, is required. Admittedly, this is only useful if F can be learned, and if the space is not too spiky (nearby values for w should, in general, imply nearby values for $M(E(u))$). Nevertheless, the the third strategy would be useful for situations in which immediate, gist-based action is required.

E. Metamappings as metacognition

As explained at the beginning of this section, we have taken these efforts to incrementally motivate the architecture in strategy three in order to illustrate how a grounded, sensorimotor based system can merit ascription of the kinds of metacognitive abilities that we have proposed are necessary for crediting a system with qualia:

- In effect, the forward model E confers on the the system belief-like states, in the form of expectations of what sensor values will result from performing a given action. These (object, not meta) belief-like states are total in that a given state vector w yields an E_w that manifests a range of such expectational beliefs, each concerning a different action or point of fixation.
- Similarly, the forward model F confers on the the system meta-belief-like states, in that they indicate which total, object belief states have a particular content property. (Note that the meta beliefs are not of the form, for some particular w, u and cue: w manifests the belief that (or represents that) $M(R(u)) = M(cue)$. Rather, they are of the form, for some particular w and cue: $\exists u : w$ manifests the belief that $M(R(u)) = M(cue)$.)

Meta-belief is not only an explicit requirement for the kind of qualia-supporting architecture outlined in section II and III; it also opens to door to the further requirements of inference, deliberation and sensitivity to logical relations. To see how, consider one more addition to the architecture we arrived at when discussing strategy three. As with the individual nodes in the feature map, we can define the set P_{c_1,c_2} to be all those parameter sets w that yield a forward model that contains at least one expectation to see c_1 *and* one expectation to see c_2; that is, $P_{c_1,c_2} = \forall w : \exists(u_1)(u_2)$ such that:

- $M(E_w(u_1)) = c_1$; and
- $M(E_w(u_2)) = c_2$

With another network G distinct from E (and F), the robot can learn an approximation of P_{c_1,c_2}: $G(w, c_1, c_2) = 1$ iff $w \in P_{c_1,c_2}$. That is, G is a network that:

- takes the parameters w of E as input
- outputs a 1 only if those parameters realise a forward model E_w for which:
  - $\exists u_1 : M(E_w(u_1)) = c_1$; and
  - $\exists u_2 : M(E_w(u_2)) = c_2$;

Note that it is a logical truth that $w \in P_{c_1,c_2} \rightarrow w \in P_{c_1}$. It follows that there is a logical relation between G and F; specifically, it should be true that $G(w, c_1, c_2) = 1 \rightarrow F(w, c_1) = 1$. Assuming F and G are themselves reasonably accurate, the robot could observe and learn this regularity. But because F and G are only approximations, there might actually be cases (values of w) where they are inconsistent (where $G(w, c_1, c_2) = 1$ but $F(w, c_1) = 0$). That such a mismatch constitutes error could be built into the architecture, yielding an error signal not between expected and empirical object-level states of affairs, but between a logical norm and the empirical relation between meta-belief states that should respect that norm.

How should the robot respond to this error signal, which indicates the violation of a logical norm? In the case of empirical, object-level error, the direction of fit is from model to world, so error should be reduced by changing the model (*pace* Friston and active inference[7]). But in this case, the error is not between model and world, but between two models of the world: should the robot modify F, or G, or both?

Although it seems unlikely that there is a general, situation-independent answer to this question, one could certainly imagine another iteration of reflection and complexity that would enable a robot to learn an effective way for handling such situations. For example, F and G could be part of a network of experts, in which a gating network learns the kinds of situations in which any F/G mismatch should be resolved in $F's$ favour, and which in $G's$. But there is also the possibility of a resolution due to *implicit* architectural features that do not constitute a semantic ascent. An interactive activation competition between F and G might, for example, always be resolved in $F's$ favour simply because F has fewer inputs and parameters than G – or vice versa. Such a system could be understood as having a belief, albeit an implicit one, that object-level beliefs manifested in F are always more reliable, justified, etc. than beliefs manifested in G. And again, a sophisticated architecture, although continuous with the kinds of systems considered so far, could observe instances of this regularity, and thus learn the regularity itself. It could thus come to know (or at least believe) that it always takes F-based judgements to be more reliable than (logically conflicting) G-based ones. From the error signal that is produced whenever they disagree the system could come to believe that G and F are logically related. The crucial point is that the robot has the essentials of a notion of logical justification and logical consistency of its own beliefs. It could use a systematic mismatch between G and F as evidence that G requires more learning, or indeed use that mismatch as a further error signal to guide learning in G, or even E itself.

One could ask: why go to all this trouble? Couldn't all of this have been motivated simply by considering a robot that contains two forward models, E and E', that are meant to have the same functionality, but which might contingently evolve in such a way that they disagree on some inputs? The answer is yes, and no. Yes, an instance of being a logically-constrained cognizer is that one eschews believing P and $\neg P$. But no: to start with such an architecturally unmotivated example would not serve to make a general case for how meta-beliefs as a whole could get going in a sensorimotor grounded architecture. For one thing, it doesn't suggest how sensitivity to logical relations between sub-networks could assist in inference. But with what has been presented concerning the conjunctive cue network G, it is possible to understand, for example, how there could be a *disjunctive* cue network H that maps weights w to 1 only if either one or the other of its associated cues c_1 and c_2 is in the range of E_w. Such a network having output of 1 for w, in the face of $F(w, c_1) = 0$, would allow the network to infer that $F(w, c_2)$ *should* be 1, and use that in place of computing $F(w, c_2)$ explicitly, or to generate an error signal if $F(w, c_2) \neq 1$, etc.

Further sophistication, conferring even more of the kinds of metacognitive abilities discussed in sections II and III, could be added by not just allowing the robot to observe the holding or not of various logical relations in its own beliefs, but by giving it the ability to take action on the meta-level, and allow such actions to be guided, as on the object level, by expectations realized in forward models on the meta-level. Such forward models would not manifest expectations about how sensory input would be transformed by performing this or that movement, but rather how object-level forward models such as E would change, if one were to perform this or that operation on their parameter sets w. To give a trivial example, there might be a primitive operation N that could be performed on a forward model's parameters that had the effect of normalizing those parameters. A network's understanding of this might be manifested in a network J such that $J(w_1, N) = norm(w_1)$, $J(w_2, N) = norm(w_2)$, etc., with J being consulted when normalization is being considered as a possible meta-action to perform.

V. Conclusion

The "Hard core" of consciousness is meant to be qualia, but sections I-III argue that qualia, understood as the underlying phenomenon (if any) that explains qualia-talk and qualia-beliefs, might be explicable in terms of phenomena that are considered to fall under the "easy problems" of consciousness. The speculations of section IV fall short of closing the loop started in sections II and III, but they hopefully give one an idea how a grounded sensorimotor robotic cognitive architecture could merit attribution of such features as having beliefs and having beliefs about beliefs. In particular, it is hoped that some substance has been given to the possibility of such an architecture being able to employ concepts such as justifcation, deliberation and consistency.

Acknowledgment

The authors would like to thank David Booth, Simon Bowes, Simon McGregor, Jonny Lee, Matthew Jacquiery and other participants at the E-Intentionality seminar on December 1st, 2017 at the University of Sussex, and the participants at a workshop and lecture on these ideas held at the University of Vienna on December 4th and 5th, 2017, for their helpful comments on the ideas expressed in the first three sections of this paper.

References

[1] A. Sloman and R. Chrisley, "Virtual machines and consciousness," *Journal of Consciousness Studies*, vol. 10, pp. 4–5, 2003.

[2] R. Chrisley and A. Sloman, "Functionalism, revisionism, and qualia," *APA Newsletter on Philosophy and Computers*, vol. 16, pp. 2–13, 2016.

[3] D. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press, 1996.

[4] M. Tye, "Qualia," in *The Stanford Encyclopedia of Philosophy*, winter 2016 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2016.

[5] D. Dennett, "Quining qualia," in *Consciousness in Contemporary Science*, A. J. Marcel and E. Bisiach, Eds. Oxford: Oxford University Press, 1988, pp. 42–77.

[6] R. Chrisley, "Cognitive map construction and use: A parallel distributed processing approach," in *Connectionist Models: The Proceedings of the 1990 Connectionist Models Summer School*, D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton, Eds. San Mateo: Morgan Kaufmann, 1990.

[7] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty, and G. Pezzulo, "Active inference and learning," *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 862 – 879, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0149763416301336

Section 2: Short Papers