Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 2 (2020) 199–215
© World Scientific Publishing Company
DOI: 10.1142/S2705078520500101

Artificial Consciousness, Meta-Knowledge, and Physical Omniscience*

Ron Chrisley

Centre for Cognitive Science
Sackler Centre for Consciousness Science
and Department of Informatics
University of Sussex
Falmer BN1 9QJ, UK
ronc@sussex.ac.uk

Published 5 August 2020

Previous work [Chrisley & Sloman, 2016, 2017] has argued that a capacity for certain kinds of meta-knowledge is central to modeling consciousness, especially the recalcitrant aspects of qualia, in computational architectures. After a quick review of that work, this paper presents a novel objection to Frank Jackson's Knowledge Argument (KA) against physicalism, an objection in which such meta-knowledge also plays a central role. It is first shown that the KA's supposition of a person, Mary, who is physically omniscient, and yet who has not experienced seeing red, is logically inconsistent, due to the existence of epistemic blindspots for Mary. It is then shown that even if one makes the KA consistent by supposing a more limited physical omniscience for Mary, this revised argument is invalid. This demonstration is achieved via the construction of a physical fact (a recursive conditional epistemic blindspot) that Mary cannot know before she experiences seeing red for the first time, but which she can know afterward. After considering and refuting some counter-arguments, the paper closes with a discussion of the implications of this argument for machine consciousness, and vice versa.

Keywords: Knowledge Argument; Mary; Meta-knowledge; Qualia; Physicalism; Omniscience; Epistemic Blindspot; Virtual Machine; Functionalism; Illusionism; Indexical; Revisionism; Machine Consciousness; Artificial Intelligence.

1. How to Read this Paper

This paper builds on and supports prior theoretical and philosophical work by Aaron Sloman and myself on some connections between artificial intelligence and consciousness. Accordingly, the full contribution of this article is best understood in the context of our work on machine consciousness, particularly our virtual machine

[*] Sloman, A. and Chrisley, R. [2003] Virtual machines and consciousness, *J. Consciousness Stud.* **10**, 4–5; Chrisley, R. and Sloman, A. [2016] Functionalism, revisionism, and qualia, *APA Newsletter Philos. Comput.* **16**, 2–13; "Architectural requirements for consciousness," in R. Chrisley, V. Müller, Y. Sandamirskaya & M. Vincze (eds.), *EU Cognition 2016: Cognitive Robot Architectures, Vienna, Austria* (CEUR Workshop Proceedings), pp. 31–36.

functionalist account of the qualitative aspects of experience, or qualia [Sloman & Chrisley, 2003; Chrisley & Sloman, 2016, 2017]. Explicit reference to this prior work is made in Secs. 2 and 6. But the central contribution of this paper, detailed in Secs. 3 and 4, with some objections covered in Sec. 5, can be appreciated independently of its implications for machine consciousness, in that it constitutes a novel reply to Frank Jackson's Knowledge Argument (KA) against physicalism [Jackson, 1982]. Although most readers of this journal will probably wish to read this paper straight through, as written, those readers who are not interested in machine consciousness, but are only interested in the KA and its objections, can skip Secs. 2 and 6 without debilitating loss. On the other hand, familiarity with the KA is presupposed; those unfamiliar with that argument should consult Jackson's original paper or some other introductory treatment before proceeding.

The argument in Sec. 3 shows that the thought experiment Frank Jackson employs in his KA is unsound in that it asks us to suppose a situation that is logically impossible: that a person, Mary, has all physical knowledge, but that she has not experienced seeing red. This supposition is shown to be impossible by constructing a physical fact F (actually, a set of facts F) that Mary cannot know. The unknowability of F for Mary is a result of the fact that knowing F is a case of *meta-knowledge*: knowledge about someone's (in this case Mary's) epistemic situation. Section 4 presents a patch to the KA that renders it sound, but then constructs another F that renders the Revised KA invalid. This second F is a kind of meta-knowledge that Mary, despite having all physical knowledge, cannot know *if she has not experienced seeing red*. Moreover, this fact F is such that once Mary *has* seen red, it is no longer impossible for her to know it, making F a candidate for the knowledge that Mary gains when she sees red for the first time, in contradiction with the (Revised) KA. Section 5 rebuts some objections to the arguments of Secs. 3 and 4.

Without the context and independent motivation provided by Sec. 2, F (particularly its self-referential character) may seem contrived and recherché, unlikely to be the sort of thing that one comes to know when one sees red for the first time. But Sec. 2 (and later Sec. 6) links the having of qualia with dispositions to possess certain self-referential beliefs, and thus provides independent motivation for the plausibility of F playing this role, providing both a robust defense of physical qualia, and some important constraints on the design of conscious artificial agents.

2. What has Gone Before

In our previous work, Aaron Sloman and I argued that a virtual machine functionalist treatment of consciousness [Sloman & Chrisley, 2003] can make room for a nondualistic understanding of qualia [Chrisley & Sloman, 2016], which in turn suggests a concrete roadmap for an artificial intelligence-based empirical investigation into consciousness in natural and artificial systems [Chrisley & Sloman, 2017]. A key step in this prior work is an argument (drawing on causal theories of reference [Kripke, 1980; Putnam, 1975] that "qualia" refers to the physical, computational states and processes that give rise to qualia thought and talk, much like the way that "gold" refers to the physical substance that gives rise to thought and talk about gold. This understanding of the referent of "qualia" allows for many commonly-held beliefs about qualia to be false of the actual referent of "qualia", without threatening to eliminate "qualia" from scientific discourse (much like the ancients' believing many false things about gold did not eliminate "gold" (or rather its ancient cognates) from scientific discourse). This steers scientific investigation of qualia away from questions such as "how can we reconcile the immediacy, intrinsicness, ineffability and privacy of qualia with the physical world?", and toward questions such as "why is it computationally/evolutionarily/epistemically useful for conscious beings to be in experiential states that make it primitively compelling for them to believe that the qualitative character of such states is immediate, intrinsic, ineffable and private?". Questions of the latter sort can allow that qualia are real (there is something about our experience that makes us believe those things about our experience), but do not assume from the outset that qualia actually have the properties we are primitively disposed to believe them to have.

What is being advocated here is thus a form of what Dennett has called "heterophenomenology", taking people's assertions about their experience as data to be explained, without assuming that those people are correct beyond how things seem to them [Dennett, 2003]. Unlike Dennett, however, we do not take the falseness of people's beliefs about qualia to be decisive grounds for eliminating "qualia" from scientific discourse. There are also affinities here to what has recently been dubbed the "meta-problem of consciousness", which is "the problem of explaining why we think consciousness poses a hard problem, or in other terms, the problem of explaining why we think consciousness is hard to explain." [Chalmers, 2018]. Similarly, those readers who are familiar with the recently-coined term "weak illusionism" [Frankish, 2016] might find it useful to locate our position under that designation (although I prefer the term "revisionism"). Indeed, I have sometimes referred to the compulsive complex of relevant recalcitrant beliefs — that qualia have the problematic properties of intrinsicness, immediacy, ineffability and privacy — as "The Qualia Delusion".

On our account, it is a hallmark of experiences with qualia that they tend to make one believe things like "one can only know what it is like to have the experience I am having if one has actually had this experience", which in turn imply beliefs like "there is knowledge that one can only have if one has had experience X". In particular, my having an experience of red will typically involve me being in a state that makes me likely to believe "if I had never seen red, I wouldnot know what it is like to see red (that is, to have this experience)", and "if someone else (e.g. Mary) has never seen red, she cannot know what it is like to see red". Let us use F to denote the fact (or facts) that one comes to know when one learns what it is like to see red. Then not only will having the experience of seeing red make one believe F, but it will also tend to make one believe "if Mary has never seen red, then she does not know F".

3. Epistemic Blindspots: The KA is Unsound

Discussions of the KA typically construe the thought experiment it relies on as assuming that before Mary leaves the black and white room, she (a) knows every physical fact and (b) has never had the experience of seeing red.

The reason for attributing physical omniscience (a) to Mary is that it blocks a physicalist account of what Mary learns when she has the experience of seeing red for the first time. If there were some set N of physical facts that Mary did not know before leaving the room, then the fact that she learns something F when she leaves the room would not allow one to conclude that F is non-physical, since it might be one of the facts in N.

There are several reasons why one might object to a thought experiment that requires us to suppose that Mary knows all physical facts. For one thing, it could very well be that there are an infinite number of physical facts, and yet presumably Mary is finite. Similarly, the number of physical facts, even if finite, might very well require, for them to be known simultaneously, a brain many times larger than the largest possible human brain. Then there is the amount of time it would take to acquire all these facts, likely many times longer than the longest possible human lifespan.

But all of these objections (along with many others) are against the *practicalities* of actually realising the Mary scenario. As such, they allow the proponent of the KA to insist that these objections miss the point; as long as the Mary scenario is *logically* possible, they could claim, we can explore the implications of our concepts of physicalism and phenomenality by considering it. This can be disputed, of course: one could counter that the intuitions on which we rely when conducting thought experiments are less useful the more one departs from physical possibility towards mere logical possibility.

Resolving such a dispute might be difficult, and take a long time. It seems to me that a much more efficient way to challenge the KA would be to show that the premise that Mary is physically omniscient (knows all the physical facts) is not just practically unachievable, but logically impossible.

The first step to doing this is to realise that the KA makes a strong assumption: that a fact is physical implies it is knowable. Specifically, the KA assumes:

PO Mary knows all the physical facts

If it could be demonstrated that for any agent A there are physical facts F such that it is logically impossible that A knows F, then it could be shown that PO is a logical impossibility, rendering the KA unsound.

The most direct way to do this is by construction: present a proposition F such that

(a) F is true.
(b) F is knowable (e.g. by someone other than Mary).
(c) Knowledge of F is physical knowledge.
(d) It is logically impossible for F to be known by Mary.

Epistemic blindspots [Sorensen, 1984], building on [Moore, 1942] would seem to be a good place to look. An example of an epistemic blindspot is:

B "It's raining, but Ray does not know that it"s raining

If B is true, then it is an epistemic blindspot for Ray in the sense that it is something that the rest of us can know, but which Ray himself cannot. This feature of B derives from the fact that knowing B is to possess a kind of meta-knowledge (knowledge about Ray's knowledge).

We want to show that B meets the conditions enumerated above; that is

(a) B is true.
(b) B is knowable (e.g., by me).
(c) Knowledge of B is physical knowledge.
(d) It is logically impossible for B to be known by Ray.

Skip (a) for now. (b) follows from the fact that, e.g., I can know each of the conjuncts of B. (c) follows from the dialectical presumptions of the KA (if propositions such as the conjuncts of B were not physical knowledge, we would not need the KA to show us that there is non-physical knowledge).

To see why (d) is true, suppose it were false. That is, suppose there were some possible world in which Ray knows B. Suppose also a limited conjunctive closure principle for knowledge:[a]

CK If x knows P $\wedge$ Q, then x knows P and x knows Q.

Then the following reductio can be constructed in the world in which Ray knows B:

(1) Ray knows B (working hypothesis, to be rejected).
(2) B is true (from 1 and the fact that knowledge implies truth).
(3) It is raining (2, conjunction elimination).
(4) Ray does not know it is raining (2, conjunction elimination).
(5) Ray knows it is raining (1 and CK).

(4) and (5) constitute a contradiction, which establishes that the world in which Ray knows B is not a possible world after all.

So in general, (b)−(d) hold. Furthermore, in those worlds in which it is raining and Ray doesn't know it is raining, B also meets condition (a).

Could something like B serve as our F to cause problems for the KA?

Not just any epistemic blindspot will do the trick. Consider:

BM It is raining but Mary does not know that it is raining.

[a] This should not be objectionable on the part of the proponent of the KA, since in the first two sentences of "Epiphenomenal Qualia", Jackson packs closure into his definition of physical information: "It is undeniable that the physical, chemical and biological sciences have provided a great deal of information about the world we live in and about ourselves. I will use the label "physical information" for this kind of information, *and also for information that automatically comes along with it.*" [Jackson, 1982].

For reasons parallel to the above, BM meets conditions (b)−(d). But what about condition (a)? That's where the parallel ends. Given the physical omniscience assumption of the KA, Mary surely knows that it is raining. So BM is false, and is thus unsuitable for demonstrating that there is a physical fact Mary does not know.

Note also that epistemic blindspots such as BM are fragile in that, once the subject of the blindspot is informed of the truth of the blindspot, or at least its antecedent, he or she typically comes to know the antecedent, and so the blindspot becomes false.

Now consider what might be called a recursive epistemic blindspot:

R P and Mary does not know R,

where P is any physical fact. R is an epistemic blindspot for Mary, although structurally different from those considered so far in that it is explicitly self-referential. Thus, the derivation of its status as a blindspot for Mary is a little different from before. Again, using indirect proof:

(1) Mary knows R (assumption).
(2) Mary knows P (assumption of KA; also: 1, CK).
(3) Mary knows that Mary does not know R (1, CK).
(4) Mary does not know R (3, knowledge implies truth).
(5) Contradiction (4,1).
(6) Mary does not know R (1, 5, indirect proof).

Most importantly for the purposes at hand, R is true in the context of the KA, fulfilling, as BM could not, condition 1. Thus we have a proposition that demonstrates the logical impossibility of PO, the physical omniscience premise of the KA:

(1) R is true.
(2) R is knowable (e.g. by me).
(3) Knowledge of R is physical knowledge.
(4) It is logically impossible for R to be known by Mary.

Desperados might try to deny 3 on the grounds that meta-knowledge is not physical knowledge, but as remarked before, this move backfires against the proponent of the KA: if facts fail to be physical simply because they are meta-knowledge, then we hardly need the KA to establish that there are non-physical facts; to assume that such facts are non-physical is to beg the question.

The upshot is that the KA as originally put forward by Jackson is unsound, and fails.

4. Recursive Conditional Epistemic Blindspots: The Revised KA is Invalid

Admittedly, R seems contrived; the problems it raises for the KA seem like a mere technicality. One might wonder if there is a simple patch to the KA that can avoid these problems while retaining the original power and purpose of the KA.

One proposal for revising the KA is this: in setting up the thought experiment, do not assume that Mary knows all the physical facts (since we now know that to be logically impossible); instead assume that Mary has *limited* physical omniscience:

LPO Mary knows all the physical facts that can be known (by Mary).

This will leave a residue of facts N, that are physical but unknowable by Mary. But if we grant Jackson his intended outcome of the thought experiment (that when Mary leaves the room and sees red for the first time, she learns a new fact F), the case against physicalism can proceed much as before. In order for Mary to learn F, it must be a fact she did not already know before leaving the room. Although there are physical facts N that Mary did not know before leaving the room, none of these can be F, since the facts in N are unknowable by Mary, yet Mary comes to know F. So F cannot be physical and knowable by Mary, nor can it be physical and unknowable by Mary. Therefore, it cannot be a physical fact at all. Thus, there are non-physical facts. The revised KA (the version employing LPO) stands.

Although R is useful in demonstrating that N is non-empty, it fails to be of assistance in attacking the revised KA directly, in that it cannot play the role of F. R is never knowable by Mary, whereas F is known by Mary after she leaves the room.

A key insight for defeating the revised KA along lines similar to those in Sec. 3 is this: whether or not a fact is knowable by Mary can change, depending on what else is true.

Consider the following schematic proposition:

RX P and (if X has not seen red, X does not know RX),

where P is a physiological fact or conjunction of physiological facts having to do with the processing of red light by the eye and brain. (Similarly to R, RX is a recursive *conditional* epistemic blindspot because it is a *conditional* epistemic blindspot [Sorensen, 1984] that refers to itself.) Now, consider a specific instance of RX:

RM P and (if Mary has not seen red, Mary does not know RM).

To know RM is possess meta-knowledge about Mary's epistemic state. RM is an epistemic blindspot for Mary *if she has not seen red.* To see why, suppose Mary has not seen red, but she knows RM:

(1) Mary has not seen red (hypothesis).
(2) Mary knows RM (working hypothesis, to be rejected).

To proceed further, we require two auxiliary assumptions. The first is

NR Mary knows that she has not seen red.

One could try to derive NR from the fact that Mary has not seen red, together with some negative transparency thesis, such as

NT If x has not has the experience of seeing E, then x knows x has not had the experience of seeing E.

But NT seems too strong in general. Consider Jane who, like Mary, has never had the experience of seeing color, but is not physically omniscient. In fact, Jane does not even know that there are such things as colors, and has never heard of red. Despite the fact that she has never had the experience of seeing red, it seems wrong to attribute to Jane the knowledge that she has not had the experience of seeing red. So NT is false.

Fortunately, NT is not needed to establish NR. Presumably, the knowledge NR attributes to Mary is physical knowledge; if it is not, we do not need the KA to refute physicalism. We also have no reason to believe that the fact that Mary has not seen red is unknowable to Mary. So NR follows from Mary's limited physical omniscience (LPO).

Next, we need to assume another reasonable (for the same reasons given for CK) closure principle for knowledge:

CK2 If x knows (P → Q) and x knows P, then x knows Q.

Our proof that RM is a blindspot for Mary continues:

(3) Mary knows (if Mary has not seen red, Mary does not know RM) (2, CK).
(4) Mary knows that she does not know RM (3, NR, CK2).
(5) Mary does not know RM (4, knowledge implies truth).
(6) Mary does not know RM (5, 2, indirect proof).
(7) If Mary has not seen red, Mary does not know RM (1, 6, conditional proof).

The upshot of the fact that RM is a (conditional) epistemic blindspot for Mary is that the revised KA, with its assumption of LPO, also fails. To see how, note that in the context of Mary not having had the experience of seeing red, RM is true, and is knowable by others (in fact, the forgoing reasoning that establishes RM as a (conditional) blindspot for Mary ipso facto gives anyone other than Mary knowledge-granting grounds for believing RM to be true, assuming they already have grounds to believe the first consequent of RM). So in this context, RM is in N, the set of physical facts unknowable by Mary. So LPO does not suppose that Mary knows RM before having had the experience of seeing red.

But notice what happens to RM once Mary *has* had the experience of seeing red. RM is still true, since the first conjunct remains true, and the second conjunct, a material conditional, has a false antecedent and is therefore also true. And of course RM remains a physical fact. But RM is *no longer* an epistemic blindspot for Mary: the upshot of the falseness of the antecedent of the second consequent of RM is that RM can be true without requiring the truth of the consequent of the second conjunct of RM. That is, there is no contradiction in supposing that Mary now knows RM.

Thus, RM is a physical fact that can play the role of F: something that Mary does not know before having the experience of seeing red, but which she can know after having had the experience of seeing red. And there are an unbounded number of physical facts like RM that can play this role: simply replace the first conjunct of RM with any another physical fact or conjunction of facts, in particular facts concerning color vision. Consider also RXI, which is RX instantiated with the first-person indexical, rather than third-personal references to Mary:

RXI P and (if I have not seen red, then I do not know RXI).

Consider also that P may contain quantified or instantiated versions of RX applied to others:

P = P1 and P2 and... and (For all conscious subjects Y, if Y has not seen red, then Y does not know RY) and...

The revised KA, employing LPO, cannot rule out the possibility that it is one of these physical facts that Mary learns when she has the experience of seeing red.

Thus the revised KA (using the limited physical omniscience assumption), like the unrevised KA (using the unlimited physical omniscience assumption), fails.

5. Objections

5.1. *Objection: Infinite facts?*

Consider the expansion of RM:

RM P and if Mary has not seen red, then Mary does not know {P and if Mary has not seen red, then Mary does not know {P and if Mary has not seen red, then Mary does not know {P and if Mary has not seen red, then Mary does not know {P and if ...}.

This expansion is clearly infinite. There are two problems with this:

- (C1) No physical facts can be infinite.
- (C2) Even if C1 is false, no infinite physical fact can be known by a finite physical system, so while RM may be in N it cannot ever move from N to what Mary knows (undermining the argument against the revised KA).

My reply: If C1 is correct then RM, while a fact, is not physical, since it is (let us grant) infinite. So we already have a refutation of physicalism; we hardly need the KA. But less flippantly, consider that C2 can be refuted directly by the fact that *you* know RM, because you know P (we are assuming), and you know that Mary does not know RM (because you can see that a contradiction would ensue if she did). And yet you are a finite being. So, can it really be the case that RM is infinite in any physically problematic sense?

I should make clear that the exotic, self-referential aspect of RM that some find suspicious is essential to its suitability for playing the role of F. For example, consider the non-self-referential proposition:

NM If Mary has not seen red, then Mary does not know what it is like to see red.

Even though this yields a weak form of self-reference when Mary believes it, it is not enough to make it an epistemic blindspot for her: there is no contradiction in supposing that Mary knows NM, regardless of whether she has or has not seen red. Therefore, it is unsuitable to play the role of F. In fact, anyone who has grasped (or who is in the grip of!) the concept of qualia plausibly believes something like NM of themselves, and for many or all kinds of experience, even (or especially) for those experiences which they have not yet had (assuming that knowledge of such experience is not (believed to be) constructible from experiences they have already had — cf Hume's "missing shade of blue").

5.2. *Objection: P is superfluous* 1

Why do you need to have 'P' in your 'R' facts? Why not just use:

RM' If Mary has not seen red, she does not know RM'

as your candidate for F that is both a physical fact and something that Mary can come to know?

My reply: Although RM' can play this role, I think it is best to see RM' as a special (degenerate) case of the schema of RM. Doing so reveals that there are an unbounded number of physical facts that could play the role of F, not just one. If someone had an argument why RM' could not be what Mary learns (and it does, on the face of it, seem implausible), this refutation of the KA would fail if I only relied on RM'. But by using RM, it becomes clear that there might very well be a very particular physical fact, out of the infinitely many facts that can play the role of F, that Mary actually learns when she sees red for the first time.

5.3. *Objection: P is superfluous* 2

> Pressing on with the previous objection: Is it not implausible that RM could be what Mary learns when she experiences seeing red for the first time? By assumption, she knows everything in the first conjunct already, before seeing red for the first time; how can just adding the second, self-referential conjunct to these already-known facts yield a fact which is what Mary comes to know when she sees red for the first time?

The first, modest point to make here is that RM does not need to be plausible in order to establish that (a) the original KA is unsound and (b) the revised KA is invalid. RM, as it stands, is only meant to serve as a counter-example to the conclusions usually drawn from Jackson's thought experiment. To show that the KA fails, I do not have to produce the actual fact F that Mary would learn when she sees red for the first time; I only need to show that the argument that F cannot be physical fails. Which I believe I have done.

But a more robust reply can be made by refusing to conflate two things: on the one hand, what would be learned by hypothetical, physically omniscient Mary if she were to experience seeing red for the first time, and on the other hand, what would typically be learned by an actual, non-physically-omniscient person, Jane, upon seeing red for the first time. It seems plausible to suppose that even if Mary learns nothing except RM (or even RM'), that is consistent with Jane learning many other physical facts, in addition to RM or RM' (or rather their equivalents, RJ and RJ', that mention Jane instead of Mary). We should not be tempted to infer from the richness of what Jane would learn to a similar richness in what Mary would learn. The sparseness, if any, in RM may simply be down to the fact that by hypothesis Mary already knows so much; she already knows everything *else* about what it is like to see red — all she lacks is the physical knowledge that she is logically prohibited from having.

This is distinct from, but motivated by, much of what Daniel Dennett has said on the subject. So that I might defeat the strongest form of the KA, I have been conceding the intuition Jackson asks us to share, that Mary will learn something F upon seeing red for the first time. Indeed, much of my reply consists in showing that there is a physical fact that can play role of F. But Dennett gives good reasons for questioning this intuition, arguing instead that Mary will not learn anything at all, if she really is physically omniscient. Failing to recognize this, he says, is a failure of imagination, relying too much on our intuitions about how normal, non-omniscient people would fare in such a bizarre situation. Note that if Mary is physically omniscient (and particularly if that omniscience implies continued physical omniscience after leaving the room), then there will be no *practical* ability or behavior concerning color that Mary will lack before seeing red that she will gain upon seeing red for the first time (and the same, of course, goes for yellow, or blue). For example, despite the intuitions some may have, Mary will not be fooled upon being presented with a blue banana into thinking "oh, so this is what yellow looks like!", since her continued physical omniscience will tell her that the activity in her visual cortex is characteristic of seeing blue things, not yellow things [Dennett, 1992, pp 399–400].

I am inclined to agree with this much, but then Dennett takes things just a little too far, by asserting that if Mary has all (logically possible) physical knowledge, then she will learn nothing upon seeing red for the first time [Dennett, 2006]. As shown in Sec. 4, this is not technically correct, since, for example, Mary will learn RM. But what is the *practical* upshot of learning RM? For a physically omniscient Mary, the answer is almost certainly: nothing. This does not mean, however, that non-physically-omniscient Jane learns nothing practical when *she* experiences seeing red for the first time. Her version of RM, RJ, will undoubtedly contain a lot of knowledge, previously unknown to Jane, in its version of P. So, this is another reason why it is useful to use RM rather than RM' to express what is learned upon seeing red for the first time: the P allows us to pack in less or more physical knowledge to be learned depending on the amount of physical knowledge already possessed by the person in question (e.g. omniscient Mary versus relatively ignorant Jane).

5.4. *Objection: RM is too easy to know*

> While it is true that Mary cannot know RM unless she has seen red, it is not a good candidate for F (what Mary learns when she sees red for the

> first time) since anyone other than Mary *can* know RM, *even if they have never seen red.*

Well, not just *anyone*; only those who know the facts in P (see the preceding subsections for why P is not superflous). Even so, some might think the objection still has a point. Consider Anne, who is in the same state of (limited) physical omniscience as Mary is before she leaves the room. Anne knows P. So, Anne knows RM. This seems odd, since RM is what Mary learns when Mary sees red for the first time, but presumably Anne would still learns something new if Anne were to see red for the first time.

It is possible to make a logically austere response to this objection, dismissing it as irrelevant: RM still provides a counter-example to the Revised KA. And it is clear how to construct a similar fact, RA, that can play the role of F for Anne. End of story.

But we can do better than that. The intuition we need to do justice to is this: for two agents X and Y of equivalent epistemic status, there is something in common between the fact that X learns when X sees red for the first time, and the fact that Y learns when Y sees red for the same time. In fact, the intuition is that, in some sense, what X and Y learn is the *same* fact. So, if X does not know that fact before seeing red for the first time, Y should not know *that same fact* before seeing red for the first time. As things stand, this is not so: before seeing red, Anne knows RM, and Mary knows RA.

One could try to accommodate this by universally quantifying the second conjunct of RM, so that it applies to everyone:

RQ P and (for all subjects x, if x has not experienced seeing red, then x does not know RQ).

But this raises tricky questions of the conditions under which one knows a universally quantified sentence. Anne is unable to know one instance of the second conjunct of RQ: "If Anne has not experienced seeing red, then Anne does not know RQ". And this is enough to bar her (or anyone else who has not seen red) from knowing the universally quantified portion of RQ, and thus RQ itself. But she does know (presumably) all the other instances of that second conjunct. So, it seems misleading to say that Anne learns the universally quantified statement upon seeing red, when all she really learns is the instance involving herself. Which puts us back to square one.

A better way to accommodate the intuition that, in a sense, (limited physically omniscient) Mary and Anne learn the same thing upon seeing red for the first time is to express RM (or rather, RX, where things started) in terms of an indexical:

RI P and (if I have not experienced seeing red, then I do not know RI).

This, then, is the same piece of knowledge that Mary and Anne are both missing before, and the same piece of knowledge that they both acquire after, they experience seeing red for the first time. So, RI can serve as F for both of them. RI, unlike RM, is not too easy to know.

5.5. *Objection: RM (or RI) is too hard to know*

> While it is true that Mary cannot know RM unless she has seen red, it's not a good candidate for F (what Mary learns when she sees red for the first time), since there are conscious subjects who can come to know what it is like to see red who cannot know F. Consider Jane who lacks the capacity to possess meta-knowledge (perhaps because she lacks concepts of knowledge or belief). Jane therefore cannot come to know RI. But it seems she can nevertheless come to know what it is like to see red: suppose she, like Mary, had spent her life in a black and white room, had never experienced seeing red, etc. but eventually escapes the room and sees a red apple. Surely she comes to know something thereby: what it is like to see red. And surely what Mary comes to know when she experiences seeing red for the first time must have something in common with what Jane comes to know when she experiences seeing red for the first time. But it cannot be RI, because Jane cannot come to know RI. So, RI cannot be what *Mary* learns when she experiences seeing red for the first time.

This objection makes a several questionable assumptions. It assumes that a subject can be conscious without being able to think of knowledge, belief, etc. even in an implicit, or non-conceptual sense. But even if we grant that assumption, we are also asked to assume that a subject without even implicit, non-conceptual ways of thinking of knowledge and belief can come to know *what it is like to see red.* To see how tall an order this is, consider: it is one thing to see red. It may be another thing to experience seeing red. But it seems to be yet a third thing to *know what it is like to experience seeing red.* And to me it seems at least possible, if not likely, that a subject that does not warrant ascribing to them even an implicit or non-conceptual notion of knowledge or belief will equally fail to warrant ascribing to them knowledge of what it is like to see red — merely ascribing to them the experience of seeing red will do. Why? Because ascription of the more sophisticated kind of knowledge will only be justified for a subject that is capable of comparing experiences, as individuated by what those experiences are (typically) *of.* And such capability requires at least a non-conceptual notion of mental states of a subject that (a) have a mind-to-world direction of fit, and (b) have a normative connection with behavior (such as inference, or reports of one's experiential state). And this would be enough, it seems to me, to be able to believe (and thus know, in the right circumstances), a non-conceptual version of RI.

But for those who are not willing to join me out on that limb, there is another response: First, note that we are primarily concerned here with refuting the (Revised) KA, and thus primarily concerned with the existence of a fact that can play the role of F *for Mary.* By hypothesis, Mary must be capable of having meta-knowledge, since (at least some of) such knowledge is undeniably physical (else we would not need the KA), and Mary possesses all physical knowledge. So, knowing RI is not "too hard" for *her.* Second, following the Dennett-esque line offered before, nearly all of what a normal metacognitive subject learns when they experience seeing red for the first time is "object level" physical knowledge, not meta-knowledge. But some of it is meta-knowledge (e.g. RI). Denote the non-metaphysical (not to be confused with non-metaphysical!) knowledge a typical non-omniscient subject acquires upon experiencing seeing red for the first time as NM. Then the position is: Upon experiencing seeing red for the first time, Mary learns only RI, non-physically-omniscient metacognitive subjects learn RI and NM, and non-metacognitive subjects learn only NM. That is all the commonality that the argument against the revised KA requires.[b]

5.6. *Compatibility with other objections to the KA: Modes of Presentation*

In identifying these flaws in the KA and Revised KA, I do not mean to suggest that all other criticisms of the KA are mistaken. For example, some have argued that Mary does learn something new upon seeing red for the first time, but what she learns is not a new (non-physical) fact, but an old (physical) fact under a new guise, or mode of presentation — possibly indexical [Lycan, 1996]. Even if this is correct, it is not in tension with the criticisms of the KA and Revised KA given here. Such facts could be included, under the correct mode of presentation, in the P component of RM. This "mode of presentation" reply to the KA has itself been objected to [e.g. Mandik, 2010], on the grounds that physically omniscient Mary would already know all the identities between modes of presentation, so she would not learn anything new when confronted with a particular mode of presentation of an old physical fact upon release from the room. But, I argue, this is mistaken: yes, Mary might know all the mode of presentation identities, but she might not know *the identities* under all modes of presentation (e.g. she might know them only under pleonastic modes of presentation), thus preserving the informativeness of the new mode of presentation under which Mary learns F upon seeing red for the first time. But there is no space here to explicate this counter-argument further.

6. The Upshot for Artificial Intelligence and Consciousness

The connection between research into artificial consciousness on the one hand, and the arguments and objections of Secs. 3–5 on the other, is bi-directional: what the former offers the latter, and vice versa.

6.1. *What machine consciousness research offers the Recursive epistemic blindspot argument*

Some people may accept the arguments offered in Secs. 3 and 4 at face value, taking them to be convincing refutations of the original KA and the Revised KA. Others, however, may be put off by what they perceive to be the contrived and recherché

[b] The focus on knowledge throughout this paper is a consequence of the KA's focus on knowledge. It should be stressed that the position put forward here does not imply that having a new experience exhaustively consists in the acquisition of new knowledge.

character of, e.g. RI. Is this some kind of linguistic trick that merely calls Jackson out on a technicality? Could knowledge of such an odd, recursively self-referential, indexical, conditional epistemic blindspot have any substantive connection to what it is like to experience red?

In previous work, cited at the beginning of this essay, Aaron Sloman and I have given reasons to believe that if the term "qualia" refers to anything at all, it refers to architectural features of an agent that (partly) explain why the agent has certain kinds of beliefs about (the qualities of) its own experiences (such as, following [Dennett, 1988], that they are private, intrinsic, immediate or ineffable). It follows from our account (constructed over multiple decades, completely independently of these current considerations of the KA) that when an agent A with such an architecture sees red for the first time, A acquires (possibly in addition to some non-self-referential beliefs) the (self-referential) belief that A itself has knowledge K now that A could not have before seeing red for the first time. Such a belief can be proved by A to be true, as above, making it *meta-knowledge*.

Another component of this prior, independently-motivated work is that states with *causal indexicality* [Campbell, 1994] are important to understanding the architectural basis of consciousness and qualia. This comports well, at least in general terms, with the indexicality of RI, though the exact relation between the two demands fuller discussion.

Thus, this previous work should go some way toward alleviating the doubts that some might have about the actual relevance, beyond being a rebuttal to the Revised KA, of meta-knowledge such as RI to understanding the process of coming to know what it is like to have a new experience.

6.2. *What the recursive epistemic blindspot argument offers machine consciousness research*

To start with the weakest point: any refutation of an argument against dualism is, presumably, to the benefit of machine consciousness, since the latter, one might think, requires physicalism (or at the very least, does not sit well with dualism). Put another way: one might have taken the KA to establish that artificial consciousness is impossible; if so, the arguments provided put machine consciousness back on the table of possibilities. Even if true, the weakness of this point lies in the fact that it also applies to any other (successful) reply to the KA (or indeed to any other successful reply to any other argument against dualism).[c] Another weakness of the point is that there is nothing in the arguments against the KA offered here that is specific to machine consciousness.

A stronger and more practical connection is this: the arguments offered in Secs. 3 and 4 follow the KA's lead in focussing on the connection between consciousness and

[c] Worse, it is not clear that the weak point is strictly true. Conventional wisdom does have it that the possibility of machine consciousness demands physicalism, but does it really? Dualism is, in principle, just as compatible with machines being conscious as it is with humans being conscious, it seems to me. But there is no space here to discuss this issue.

metacognition, but suggest constructive questions in the context of artificial consciousness: what connections between conscious states, and knowledge of what it is like to be in those conscious states, are required for machine consciousness? What generates those connections (evolution, computational efficiency, communication, logic?) How can those connections be supported by an machine consciousness architecture? How should we think of systems that fail to respect these connections, but otherwise meet the conditions for the attribution of consciousness? How does what a subject learns upon having a new kind of experience vary with the knowledge that subject already has?

Some computational architectures better facilitate the modelling of meta-knowledge than others. In most symbolic architectures adding the capacity for meta-knowledge would require merely adding the KNOWS (x, y) relation to an agent's stock of concepts. But matters are quite different for the grounded, bottom-up, sub-symbolic, non-conceptual architectures that are typical of machine learning-(especially neural network-) based cognitive architectures. There, it can be difficult to talk of knowledge or concepts at all, let alone meta-knowledge, which requires possession of the concept of knowledge. Such architectures otherwise have great promise for modeling consciousness; do the connections between meta-knowledge and consciousness made here and in the earlier work cited at the beginning of this paper exclude the use of machine learning architectures for machine consciousness? I do not believe so; I give some reasons for optimism, and make some positive architectural suggestions along these lines, in [Chrisley, 2018].

In closing, it must be admitted that the restrictions at the centre of the arguments — e.g. that a conscious agent (natural or artificial) that has not had the experience of seeing red cannot know RI — are logical: an artificial agent will meet them automatically. Compare: one does not have to design a robot so that it respects the restriction that an object cannot be in two places at once; that comes for free. Not being logically permitted to know RI if one has not had the experience of seeing red, and *being logically permitted* to know RI if one has, comes for free too. But there is an enormous gap between logical permissibility and actuality. What does not come for free, and thus must be designed for in some more proactive sense, is actually knowing RI in the situations in which it is logically permissible to do so. If this objection to the (Revised) KA is correct, moving beyond the mere possibility of knowing RI to actually knowing it may be an important, even necessary part of a metacognitive agent's coming to know what it is like to have a new kind of experience. This provides a guiding constraint on architectures for machine consciousness in metacognitive agents.

Acknowledgments

This paper was completed while I was a Visiting Professor in Symbolic Systems at the Institute for Human-Centered Artificial Intelligence at Stanford University. The Fellowship that supported my stay was made possible by the University of Sussex and generous donations from Reid Hoffman and the Aphorism Foundation. I benefitted from discussions of these matters with several people, in particular Matthew Jacquiery (responsible for the objections in Sec. 5.1), and Oliver Sharpe and Brian Cantwell Smith (each independently responsible for the objections in Secs. 5.2 and 5.3). Thanks also to Aaron Sloman, and to Tim Hawks for helpful discussions.

References

Campbell, J. [1994] *Past, Space and Self* (MIT Press, Cambridge).

Chalmers, D. J. [2018] The meta-problem of consciousness, *J. Conscious. Stud.* **25**(9–10), 6–61.

Chrisley, R. [2018] "Grounded metacognitive architectures for machine consciousness," in A. Chella, D. Gamez, P. Lincoln, R. Manzotti & J. D. Pfautz (eds.), *Proc. Papers of the 2019 Towards Conscious AI Systems Symposium co-located with the Association for the Adancement of Artificial Intelligence 2019 Spring Symposium Series (AAAI SSS-19), Stanford, CA, 2019* (CEUR Workshop Proceedings 2287).

Chrisley, R. and Sloman, A. [2016] Functionalism, revisionism, and qualia, *APA Newslett. Philos. Comput.* **16**, 2–13.

Chrisley, R. and Sloman, A. [2017] "Architectural requirements for consciousness," in R. Chrisley, V. M'uller, Y. Sandamirskaya & M. Vincze (eds.), *EUCognition 2016: Cognitive Robot Architectures, Vienna, Austria*, (CEUR Workshop Proceedings), pp. 31–36.

Dennett, D. [1992] *Consciousness Explained* (MIT Press, Cambridge).

Dennett, D. [2003] Who's on first? Heterophenomenology explained, *J. Conscious. Stud.* **10**, 19–30.

Dennett, D. C. [1988] Quining qualia, in A. Marcel & E. Bisiach (eds.), *Consciousness in Modern Science* (Oxford University Press).

Dennett, D. C. [2006] What robomary knows, in T. Alter & S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism* (Oxford University Press).

Frankish, K. [2016] Illusionism as a theory of consciousness, *J. Conscious. Stud.* **23**(11–12), 11–39.

Jackson, F. [1982] Epiphenomenal Qualia, *Philos. Quart.* **32**(127), 127–136, doi: 10.2307/2960077, https://doi.org/10.2307/2960077.

Kripke, S. [1980] *Naming and Necessity* (Harvard University Press).

Lycan, W. [1996] *Consciousness and Experience* (MIT Press).

Mandik, P. [2010] Mental representation and the subjectivity of consciousness, *Philos. Psychol.* **14**(2), 179–202.

Moore, G. E. [1942] A reply to my critics, in P. A. Schilpp (ed.), *The Philosophy of G. E. Moore* (Open Court).

Putnam, H. [1975] The meaning of meaning, in H. Putnam (ed.), *Mind, Language and Reality: Philosophical Papers*, Vol. 2 (Cambridge University Press, Cambridge).

Sloman, A. and Chrisley, R. [2003] Virtual machines and consciousness, *J. Conscious. Stud.* **10**, 4–5.

Sorensen, R. [1984] Conditional blindspots and the knowledge squeeze: a solution to the prediction paradox, *Austral. J. Phil.* **62**, 126–135.