Computational theories of conscious experience: between a rock and a hard place

Gary Bartlett

1. Introduction

The computational theory of mind (CTM) has drawn a familiar class of counterexamples, such as John Searle's assemblage of beer cans and string (1980), or Ned Block's Chinese nation (1978). These weird devices mimic the computational structure of conscious beings, so CTM entails that they could have experiences. Searle, Block and others take this consequence to show that CTM must be false.

Like Tim Maudlin (1989), who calls these cases *ploys of funny instantiation*, I am unimpressed. Computationalists should – and often do – bite the bullet and say that such systems would indeed have experiences (e.g., Lycan 1987).

Ploys of funny instantiation often target computationalism about *conscious experience*, but they apply just as well – or just as unsuccessfully! – to CTM at large. By contrast, I shall pursue a problem that specifically targets the computational theory of conscious experience (CTE).

Here is a sketch of what follows. In reply to the claim that mundane objects such as rocks can mimic the computational activity of conscious beings (by implementing a certain computation), computationalists say that rocks do not compute (and so cannot instantiate mentality) because they lack the requisite input sensitivity. This *input sensitivity requirement* is plausible for cognition, but much less so for conscious experience. Maudlin (1989) and Mark Bishop (2002a, b) argue that the requirement entails that two systems can differ in their experiences just by differing in some inactive physical structure. Colin Klein (2008) has tried to help CTE avoid this entailment. I argue that he fails, and generally that if CTE is to avoid panpsychism, it must deny that experience supervenes on physical activity. But as Klein himself accepts, and as I briefly discuss, there are very good reasons to think that experience *does* supervene on physical activity. Therefore, there can be no (non-panpsychist) computational account of experience.

2. The panpsychist threat: ploys of wild instantiation

2.1. The cases. Bill Lycan (1987) describes a case offered by Ian Hinckfuss (at a 1978 conference, says Copeland 1996). 'Hinck's pail' is a translucent pail of spring water left in the sun. It is full of activity: convection currents, multiplying microbes, and so on. The activity, Hinckfuss observed, might briefly implement a computation, with certain physical impingements on the pail (e.g., sunlight) counting as input. So if implementing some computation *c* suffices for having some mental state *m*, then the pail could very well have *m* by implementing *c*.

Two later arguments extend the idea to outwardly 'inert' objects – which still, of course, contain great activity at the molecular level. Hilary Putnam offers a detailed argument for the claim that "Every ordinary open system is a realization of every abstract finite automaton" (1988, p. 121). He holds that an object implements a given finite-state automaton (FSA) over a time if the object's physical states and physical state transitions over that time can be mapped onto the FSA's formal states and formal state transitions, and if each physical state causes the transition to the next. Such mappings, he argues, are freely available. He does not make an example of any particular object, but as David Chalmers (1996) suggests, a rock would do nicely. In a similar (but less detailed) vein, John Searle argues that his office wall could be "implementing any program, including any program implemented in the brain" (1992, p. 209).

Expanding Maudlin's whimsical terminology, I call such cases *ploys of wild instantiation*. If they succeed, then CTM (not just CTE – these cases look equally threatening concerning consciousness or just mentality in general) entails that not only do mundane objects implement computations, but they implement very many computations all at once – '*pan*pancomputationalism' as Piccinini (2008, p. 60) calls it. Thus they may also realize many mental events all at once. CTM is thus threatened with an especially virulent form of panpsychism: *pan*panpsychism. If computationalists cannot defuse this threat, then (assuming that they cannot accept panpanpsychism) they will have to admit that computation is not sufficient for mentality – which is effectively to give up on a computational theory of mind.

2.2. The standard response. Even Lycan (1987), as merry a bullet-biter as they come, admits that the threat of panpsychism recommends against biting the bullet on wild instantiations. The computationalist needs to explain why wild instantiations could *not* be conscious.

The explanation Lycan suggests, and which is fleshed out in detail by Ron Chrisley (1994), David Chalmers (1996), and Jack Copeland (1996), is as follows. A rock (e.g.) mimics events in a computer only in an extremely limited way. Some activity in the rock may happen to mimic the activity of a computer in response to some particular input. But had the input been different, the rock would *not* have mimicked the computer's activity in response to *that* input. And a plausible condition on an object's being a computer is that it can respond to different input in appropriately different ways. The rock lacks this input sensitivity. It merely mimics a single execution episode in response to a single set of input.

So computationalists enforce an *input sensitivity requirement* (ISR) according to which a system is computing only if it satisfies a range of counterfactuals dictating appropriate state transitions in response to a range of input. Hinck's pail, Putnam's rock and Searle's wall fail the ISR.

The ISR nicely fits what we would expect of a theory of mind. As Chalmers (1996) points out, cognition is not merely a clockwork march through a set sequence of states. Cognition requires that different input must be able to stimulate different activity. Thus the ISR seems like an ideal reply to ploys of wild instantiation.

However, trouble lurks for computationalists who try to apply the ISR to their theory of conscious experience. While the ISR is a plausible requirement on cognition, it is less plausible for experiential states.

3. The problem of experience and activity: ploys of episodic instantiation

3.1. Bishop. Mark Bishop (2002a, b) argues that failing the ISR cannot make a system fail to have experiences. He notes that once a system's input is specified, its state structure effectively reduces to a single execution episode. An execution on that input needs to implement only the

formal states *in that episode*. The rest of the formal structure is unused, so its implementation must be irrelevant to any resulting experiences.

To illustrate, Bishop imagines a physical system R_1 that, given input (I), engages in an episode of physical activity that implements a computation that, in turn, realizes an experience of a red square. Then he deletes all the state transition sequences that R_1 *does not enter* during its run on (I), by deleting the physical parts that are not called upon during that episode of activity. The result is R_2, a truncated system that can compute *only* on (I), for it lacks the structure to deal with other input.[1] R_2 does not meet the ISR, for its physical structure does not satisfy counterfactuals about any state transitions except those needed for (I). So, says Bishop, although R_1 and R_2 run on (I) with physically indistinguishable episodes of activity, computationalists are committed to the absurd claim that R_2 does *not* experience the red square.

With terminological inspiration again from Maudlin (1989), and also Klein (2008), I call such cases *ploys of episodic instantiation*. While a wild instantiation mimics computational activity only by chance, an episodic instantiation engages in such activity because it is a fragment of a legitimate computer. There is therefore a strong intuition that an episodic instantiation must be experientially the same as the computer of which it is a fragment. Yet it fails the ISR.

Some may worry that I am relying on an unexplicated notion of 'physical activity'. However, I do not see the notion as especially puzzling or insecure (but see §4.3 for further discussion). Roughly, two objects engage in indistinguishable episodes of physical activity when they contain intrinsically physically indistinguishable movements. Of course, no two implementations of computing machines will be *precisely* physically indistinguishable; but we need indistinguishability to obtain only at a computationally-relevant level of description.[2]

Still, Bishop does not clarify exactly *why* R_2 must have the same *experience* as R_1. He says that the idea that R_1's experience is "contingent upon non-physical interactions with sections of its control program that are not executed [is] a form of dualism" (2002a, 376). However, it is not obvious what this claim comes to, as he does not elaborate on what he takes dualism to entail.[3] Ron Chrisley (2006) observes that if the charge of dualism were apt, then *all* computational explanations would be non-physicalist. If appeals to the counterfactuals that a system supports vio-

late physicalism, then we cannot even give a computational explanation of your laptop's behavior without violating physicalism! Chrisley takes this to show that Bishop must be wrong to condemn the ISR. He then points out that since R_1 and R_2 differ physically, there is no violation of physicalism in saying that R_1 is conscious but R_2 is not.

These matters need closer examination. Given Bishop's blunt assertion that the ISR entails dualism, Chrisley's reply is apt; but it misses the real issue, which Bishop does not elucidate. The real issue is the entailment that R_1's inactive structure contributes to its experience just by supporting counterfactuals. Whether this entailment violates physicalism depends on your account of physicalism, but we need not get into that in order to see its implausibility.

3.2. Maudlin. Maudlin (1989) got these matters concerning counterfactuals and experience into better focus, but his work has been neglected – even by Bishop himself until recently (Bishop 2009). Chalmers (1996) noted the significance of Maudlin's argument and said that it "requires an in-depth treatment in its own right" (p. 333n). However, only three authors have attempted such a treatment. Of these, only Colin Klein (2008) grasps the full import of Maudlin's case. I shall consider Klein's argument in detail in §4. I briefly address the other two replies, Barnes (1991) and Hardcastle (1993), in footnotes below.

The neglect of Maudlin's paper may be partly due to the vast baroqueness of the machine, 'Olympia', which he uses to make his case. I shall try to convey the wonder that is Olympia in as brief a compass as possible.

Maudlin begins with 'Klara', a Turing machine implementation. Instead of a tape, Klara has an infinite row of water troughs; and instead of a read/write head, a hose. An empty trough codes for a '0', a full trough for a '1'. Computation entails a series of operations on the troughs, which may be filled with water from the hose or emptied by opening a drain. Klara's machine table is set up to have her compute a function π on input τ (τ being an initial state of the troughs). This computation results in her having an experience φ.[4]

Maudlin creates Olympia by modifying Klara. He first rearranges the troughs so that each address in the run of $\pi(\tau)$ is immediately to the right of the preceding one. The hose will now move

steadily from left to right, one trough at a time. This obviates the need for a machine table, so Maudlin removes it. All that's needed is a wheeled armature to roll along the row of troughs. Water pours constantly from the hose. If an empty trough is to stay empty, a barrier is pre-installed to prevent water from entering it. If a full trough is to be emptied, the armature hits a rod that opens a drain hole. Now, so long as the troughs are pre-arranged correctly, Olympia will carry out the operations required for the computation of $\pi(\tau)$ simply by having the armature trundle along the troughs from left to right – no machine table required.

Yet Olympia is not computing π. She engages in the same activity as Klara when the input is τ, but she does it without counterfactual support. Were the input *not* τ, her activities would not accord with π. If a trough were filled when it should be empty (or vice versa), or a barrier or rod left unattached, the computation would be subverted. So far, then, Olympia fails the ISR.

Maudlin overcomes this problem by adding to Olympia an army of copies of Klara – as many as there are steps in $\pi(\tau)$. He then rigs Olympia's central unit (the one described above) so that if the input *were* to diverge from τ, the central unit would cease operation and a Klara copy would continue the computation from the point where the input diverged. Thus the counterfactuals dictating Olympia's behavior for non-τ input are now supported. She now meets the ISR.

Now here is the key point. Olympia can be made to fail the ISR again merely by removing the Klara copies. And that removal *does not affect her actual activity in running $\pi(\tau)$*.

This is a ploy of episodic instantiation much like Bishop's, but Maudlin better isolates the problem it poses for CTE. CTE cannot allow that Olympia experiences φ when the Klara copies are absent, for under those conditions she is not computing – since, as I have put it, she fails the ISR.[5] But since Klara *can* have φ, and since Klara and Olympia undergo the same physical activity on input τ, CTE must deny the thesis that "two physical systems engaged in precisely the same physical activity through a time will support the same modes of consciousness (if any) through that time" (p. 409). Such a denial is not a violation of physicalism, for a failure to supervene on physical *activity* does not entail a failure to supervene on the physical *tout court*. Rather, the problem for CTE is simply that the thesis itself is immensely compelling, even by computationalist lights, and that denying it incurs bizarre consequences.[6]

3.3. Can counterfactuals count? The best way to see the peculiarity of enforcing the ISR on experience is to see that to make an episode of activity fail the ISR, we need not *remove* any physical machinery. Just *disabling* it will do. Bishop (2002b) suggests that we damage the link between R_1's visual sensor and its frame store so that the store maintains a constant representation of a red square no matter what is presented to the sensor. It is now false that, for example, if R_1's input were a blue circle it would respond appropriately. R_1 is now insensitive to its input. Yet surely that cannot remove its experience of a red square. Maudlin's illustration is even more dramatic. We may remove Olympia's counterfactual support by suspending blocks between the gears of the Klara copies, so that if they *were* called upon, they would instantly jam and fail to operate. So according to the computationalist, the mere addition of such blocks, *not touching Olympia at all*, stops her from having an experience. This seems patently absurd.

The computationalist may reply that these manipulations do *not* falsify the relevant counterfactuals. After all, in both cases the system's global structure remains intact; it's just that an unusual situation prevents its activation. The computationalist may therefore say that the relevant counterfactuals are still true, since they will include an implicit *ceteris paribus* clause excluding outré circumstances like frame store malfunctions or blocks suspended in the machinery.

However, this reply misses the point. I grant a distinction (even if it is a vague one) between manipulations that falsify the counterfactuals and those that do not. Such a distinction seems to have explanatory use: as Chrisley (2006) says, a computational explanation of a system's behavior must advert to its ability to respond differentially to different input, and the counterfactuals it supports are obviously relevant in this regard. But this does not entail that the counterfactuals that a system supports are relevant to its experience. Put simply: experiences are occurrent states, and the mere presence of counterfactuals is not a plausible requirement on the presence of an occurrent state. The best way to drive this point home is to consider Olympia once again.

Let us grant the computationalist that suspending blocks between the gears of the Klara copies (so that non-τ input would bring Olympia to a halt) would not affect the non-τ counterfactuals, as the blocks are purely extraneous. By contrast, if we *removed* or *destroyed* the Klara copies, that would surely falsify the counterfactuals – if anything would. What about just *disconnecting* the

copies from the central (active) unit? It is not clear what the computationalist would, or should, say about that. And Maudlin suggests an even trickier case. The linkages between the central unit and the copies are metal chains. Suppose we expose the linkages to damp, so that they become rusted. There is still a mechanical connection, but if any of the copies were actually called upon, the linkages would break and the required activity would not occur. (It is easy to imagine similar cases in any system, whereby the system's global structure is intact but so shaky that actual use would ruin it.) Does *this* falsify the counterfactuals?

The point is not just that it is unclear what the computationalist should say, but rather that the very question is idle when the topic is whether the system is having an experience. Such manipulations cannot make a difference to whether an experience is produced by the central unit's activity in computing $\pi(\tau)$. Yet this is precisely what CTE commits to when it tries to hold to the ISR. Unlike Bishop (2002a), I take no stand on whether this amounts to a form of dualism. But CTE's position does seem to require belief in a very 'spooky' ability for manipulations of inert machinery to influence the machine's occurrent experiential state.

4. An attempted solution: Klein and episodic implementation

4.1. Klein's account. Colin Klein (2008) agrees that the ISR is problematic because it conflicts with the 'activity thesis' that "conscious states supervene on actual activity" (p. 143). The mere presence of counterfactual-supporting structures in our brains at a given time, he agrees, cannot affect our experiences at that time. So he offers an account of computation that replaces the ISR with a requirement that, he claims, avoids conflict with the activity thesis.

Klein proposes to develop a notion of *episodic implementation*, on which a system can compute a function on a particular input even if it cannot compute that function on any *other* input – thus removing the need for the support of counterfactuals dictating responses to such input. If conscious experiences supervene on such *episodes* of implementation rather than on *full* implementations, then Klara and Olympia (and R_1 and R_2) will have the same computational status, so the CTE may judge that they have the same experience – in line with the activity thesis.

Contrary to Bishop and Maudlin, Klein holds that R_2 and Olympia are not mere wind-up toys with no computational status. Indeed, he illustrates episodic implementation with an example like that of R_1 and R_2. A Turing machine program π has eight entries in its machine table. It is implemented on a system *S*. When *S* computes $\pi(\tau)$, only five entries get activated. Klein then introduces *S′*, whose machine table has only those five entries. *S′* can handle *only* τ, but engages in the same activities as *S* does on that input. Therefore, he says, *S′* computes $\pi(\tau)$ by episodic implementation. In general, Klein holds that ploys of episodic instantiation are legitimate computers after all: they are episodic implementations of a larger program. This makes it safe for CTE to say, in line with the activity thesis, that *S* and *S′* would produce the same experiences (if any) when run on τ.

But how does Klein intend to avoid the panpsychist threat? Won't the episodic implementation of $\pi(\tau)$ pop up everywhere in the form of wild instantiations – in Hinck's pail, for instance?

No, says Klein. He says that episodic implementations, like *S′*, differ from wild instantiations, like Hinck's pail, in their *dispositions*. Even if wild instantiations have "the appropriate dispositions at the computationally appropriate times" (p. 150), this is not sufficient for episodic implementation. The system must have "standing dispositions to respond in the correct ways that *manifest* at the correct instant" (ibid.). The difference, then, is that the wild instantiations merely *happen* to engage in the computationally-appropriate activity at the appropriate time, whereas the genuine computations engage in that activity because they are *disposed* to do so.[7]

I find this ambiguous. It is not clear exactly what differentiates wild instantiations from episodic implementations, because it is not clear what counts as a system's being *disposed*, in the relevant sense, to produce certain activity. However, I think that whatever Klein has in mind will not help CTE avoid the panpsychist threat while respecting the activity thesis.

4.2. Klein on dispositions and activity. Here is the problem I see with Klein's account. We have assumed that the physical activity of a given episodic implementation can be mimicked by a wild instantiation (such as a certain stretch of activity in Hinck's pail). Given this, the activity thesis entails that the wild instantiation and the episodic implementation must be identical in

their conscious experience. Yet Klein believes that only the episodic implementation is computing, and thus that only it can be conscious. So he must be denying the activity thesis.

Klein himself thinks otherwise. He states that "actual activity alone determines whether something is episodically implementing $\pi(\tau)$" (p. 152), and therefore that "no two things can differ in episodic implementation status without also differing in actual activity" (ibid.). Thus he holds that his account is compatible with the activity thesis.

The question is whether Klein's view is consistent. As already noted, he says that wild instantiations do not implement a computation. In order to make this stance fit with his claim that only actual activity determines implementational status, he must hold that wild instantiations always differ from episodic implementations *in their actual activity*. Or at least, he must hold that this is the case for those episodic implementations that produce consciousness.

One way to get this result would be to deny the assumption noted at the start of §4.2 above – that the physical activity of a given episodic implementation can always be mimicked by a wild instantiation. But there is no reason to doubt that such mimicry is always possible, and indeed Klein shows no sign of doubting it.[8] What, then, explains his claim that wild instantiations always differ from episodic implementations in their actual activity?

I am not sure of the answer to this question, but here is my guess. Klein highlights the presence of 'standing' dispositions in computing systems. He implies that this is what distinguishes computational activity from the merely incidental activity in a wild instantiation. He appears simply to hold that the mere lack of the relevant (standing) dispositions entails that a system's activity will not be computational, even if the very same physical activities occur. So far as I can see, the only way for this entailment to hold is if the dispositions somehow partly *constitute* the activity, such that physical activity that manifests a (standing) disposition is in some way intrinsically distinct from physical activity that is merely incidental.

Klein never announces that he takes this position, but there is some evidence that he does. The evidence is drawn entirely from two footnotes. This does not make me confident that I am interpreting him correctly; but nothing in his main text addresses the relationship between activity and dispositions.[9]

In a footnote early in his paper, Klein says that he will take 'activity', as Maudlin deploys the term, to mean just whatever property distinguishes stuff in which a computation is occurring from stuff in which no computation is occurring. This suggests a permissive reading of 'activity' which is borne out in a later footnote. In the later footnote, he is explaining that what his account says about Olympia depends on whether the Klara copies are *part* of Olympia. If they are, then the copies are "computationally active during an ordinary run of $\pi(\tau)$, because they are (part of) something that is manifesting a computationally relevant disposition" (p. 150n). So Klein is construing a system's *computational activity* at a time as including the entire state of the system at that time – even states of parts of the system that are not physically active at that time. On this construal, any interference with the Klara copies (taken as part of Olympia), even just 'blocking' them off from the central unit in the way Maudlin suggests, would change Olympia's computational activity during the run of $\pi(\tau)$. Hence Klein implies, *contra* Maudlin, that any resulting change in her experience would not violate the activity thesis.

I emphasize that Klein does not endorse the claim that the Klara copies are part of Olympia.[10] However, in order to even countenance it as an option he must hold the view that a system's dispositions partly constitute its computational activity.[11] And as I have indicated, that view seems to be the only one that allows his position to hang together. Is the view coherent?

4.3. Computational activity and physical activity. If my interpretation of him is correct, Klein wants to read 'activity', at least considered as a property of computing systems, as denoting a broader property than just physical activity.

Now it is true that not all aspects of computational activity must involve physical activity. The *absence* of physical activity, such as the absence of a current in a wire, can transmit information (Dretske 1981). If we take 'activity' in this broad sense, then *physically* inactive parts of a system may count as *computationally* active if their lack of physical activity transmits a signal.

However, the use of this idea as a response to ploys of episodic instantiation is highly dubious. It stipulates away the very distinction on which Maudlin's and Bishop's argument rests. Consider again Maudlin's statement of the activity thesis: that "two physical systems engaged in

precisely the same physical activity through a time will support the same modes of consciousness (if any) through that time" (p. 409). If one reads the term 'physical activity' as meaning 'computational activity' (however that is to be defined), then the thesis becomes simply a thesis of CTE: that conscious experience supervenes on computational activity. This is certainly not what Maudlin intended! His challenge to CTE arises precisely from the apparent fact that a system's computational status can depend on the state of system parts that are *physically* inactive. One cannot respond by simply defining away the notion of physical activity.

Indeed, such a stance threatens to undermine the very motivation for Klein's paper, which was to replace the ISR with a requirement that did not violate the activity thesis. If the only available notion of activity is one that applies to physically inactive states so long as they are deemed part of an ongoing computation, then the ISR itself may not actually violate the activity thesis – for states that support counterfactuals which dictate responses to alternate input are presumably part of an ongoing computation of π. Thus I am not sure that Klein can consistently hold that the ISR conflicts with the activity thesis but that his dispositional account does not.

Of course, someone might argue that the concept of physical activity is irrelevant to computational concerns, or perhaps too poorly defined to be useful, and that it *should* be abandoned in favor of a broader concept of computational activity. But this just seems wrong. While we do sometimes speak of the activity of a piece of machinery in a broad sense, as in locutions such as 'The air conditioner is active' – thus attributing activity to the unit as a whole – this is not the only way to speak of activity in a system. Nor is it even the most perspicuous way. When we are being careful – say, if we wish to understand how a piece of machinery works so that we can mend it – we are more precise: as in, 'The cooling unit is active but the fan isn't.' This precise approach is notably applicable to computers, given their immense functional differentiation. As I said in §3.3, physically inactive parts of a computer *can* influence its computational description, and thus can be important to explanations of its behavior *qua* computer, if they support counterfactuals that are relevant to the explanation of its current operations. But we should distinguish different *ways* in which a part of a system can have such influence. One way is by being physically active; others are by being disposed to be physically active, or by being causally related to

some part that is physically active. We can make none of these distinctions if we refuse to apply the concept of physical activity.

5. Does the activity thesis deserve respect?

Here is where we are. Computationalists propose the ISR to explain why wild instantiations (rocks, pails of water, and so on) cannot count as computing and thus cannot possess mentality. However, the ISR says that the mere presence of physically inert counterfactual-supporting physical structure can influence a system's experiences. This conflicts with the activity thesis that conscious experience supervenes on activity. Wanting to find a way to respect the activity thesis, Klein replaces the ISR with a dispositional requirement. However, his account must stipulate that the dispositions which are necessary for computation are also in some way constitutive of the very computational activity itself – such that a duplicate of the *physical* activity but which is not a manifestation of those dispositions is therefore not computational and hence not conscious. This clearly contravenes the activity thesis, which concerns physical activity and not, contrary to what appears to be Klein's reading, some broader notion of computational activity. If we want to take the thesis seriously, we must read 'activity' as referring only to *physical* activity.

So the CTE is left with little room to maneuver. If experience supervenes on physical activity, then since any physical activity that is necessary for computation can also occur in mundane objects such as rocks and pails of water, computationalists must admit that computation is universal. Yet surely they are right not to admit this (cf. Piccinini 2008).

Lest some are tempted at this point to deny that rocks and pails of water can mimic computational activity, note that any physical activity that is necessary for computation must be able to occur in a great variety of substances. Computationalists' own commitment to this tenet grounds their dismissal of ploys of funny instantiation, as noted in §1. So it is hard to see why rocks and so on would be ruled out on the grounds that they are not made of, so to speak, the right stuff.[12]

So at this point a tough-minded computationalist might say, 'So much for the activity thesis!' How might such a stance fare?

I do not think that the activity thesis is an *a priori* truth, able to be defended based on pure intuition. It needs empirical confirmation. For now, while I think there is considerable evidence for it, the evidence is more circumstantial than direct. Maudlin (1989) is too sanguine in this regard. He points to the established correlations between mental activity and brain activity, and asserts that the empirical evidence shows that experience supervenes on activity. There are two problems with this assertion. First, the evidence shows that experience *requires* activity, not that it supervenes on it. Second, the evidence concerns only *neural* activity, not physical activity in general. Maudlin finds it "not a far leap" (p. 409) from this evidence to the activity thesis, but others will say that the leap is further than it looks. Nevertheless, I shall make some remarks in defense of the thesis. I hope to show at least that any computationalist (or anyone else) who denies it will be swimming against a strong tide of not just intuition but also empirical evidence.

The activity thesis could be directly empirically tested, at least in theory. But doing so would require seeing whether the manipulation of inactive neural assemblies at a time can affect a subject's experience at that very time; and our current neuroscientific knowledge and techniques do not allow us to do this to anything like the requisite level of precision. (See below, however, for remarks on the Wada test, which is a rudimentary version of the procedure I have in mind.) The manipulation would have to involve in some way taking neural assemblies 'off-line': incapacitating them so that they would not fire even if stimulated appropriately by pre-synaptic neurons. (Antony 1994 imagines just such a set-up, in a case against functionalist theories of experience that parallels Bishop's and Maudlin's case against CTE.)[13] If subjects always reported that such manipulations had no effect on their experience, the activity thesis would be confirmed. My bet is that such tests would indeed show no effect on experience. The tough-minded computationalist will bet that at least sometimes – when the manipulated neurons had been supporting some relevant dispositions or counterfactuals – the subject would report a change in her experience.

Although we cannot do such tests at the moment, we know enough to say that neuroscience would be rocked if any such tests were to come out as the computationalist expects. It would be not unlike finding that a car's speed could be affected immediately by a brake line rupture even if the brakes were not in use at the time. Naturally, if the driver had been depressing the brakes at

the time, *then* the car's operation would be affected by the rupture; but no one would expect any effect if the brakes were not in use. Now of course, we know immeasurably less about the brain than we know about cars; but the distinction between dispositional and occurrent states seems to apply to both. So given that in your car unmanifested dispositional states do not make a difference to occurrent operations, the assumption is that the same applies in your brain: unmanifested dispositional states do not make a difference to occurrent operations, such as experiences.[14] Indeed, there is a medical test that relies on this assumption.[15] In the Wada test, an anesthetic (such as sodium amobarbitol) is injected into one of the cerebral hemispheres via the left or right carotid artery. This shuts down neural activity in that hemisphere, so that the mental and behavioral capacities of the other can be evaluated in isolation. The assumption is that during the period of testing, any experiences that the subject reports must be due entirely to the hemisphere that is still active.

The activity thesis also receives some corroboration from the commonplace observation that experiences – like occurrent mental states in general – are poised to immediately influence behavior (cf. [author's unpublished MS], in progress). Consider how seeing an upcoming traffic light turn red causes you to depress the brake, or how smelling smoke causes you to jump up in search of its source. Readiness to guide behavior is the norm for conscious experiences. Therefore, states that affect your experience at a time are also states that are maximally poised to affect your behavior at that time. Now it is plausible that inactive parts of your brain are not poised to affect your behavior in the way that active parts are, since behavior is most directly caused by *activity* in the body – muscle contractions and so on, originating in neural firings. We can again refer to the Wada test for illustration: any movements the subject makes during the test, including anything she says, are assumed to have their cause in the unanesthetized hemisphere – the same one that is assumed to be the basis for her experiences during this time. Barring a massive coincidence, the commonsense observation that experiences tend to influence behavior supports the thesis that our experiences supervene on neural activity.

None of these considerations prove conclusively that the activity thesis is true. For one thing, they concern only neural activity, not physical activity in general. But they place the weight of

the evidence very much in its favor. Thus since a computational theory of experience must deny the activity thesis, the evidence strongly implies that such a theory cannot be true.

6. Conclusion

The computational theory of conscious experience is caught between a rock and a hard place. On the one hand, ploys of wild instantiation illustrate that CTM in general, and hence CTE, cannot rest content with saying that anything that engages in a certain kind of activity counts as computing in a way that produces mentality. This condition is too easily met by ordinary objects such as rocks, walls and pails of water. There must be more to computation – at least of a kind that suffices for mentality – than a certain kind of activity. The ISR, a counterfactual requirement, is the popular choice to capture what that something more is in a way that fits nicely with how we suppose cognition to work; but concerning consciousness the ISR runs headlong into the activity thesis. Klein's dispositional account, while it jettisons the ISR, does no better in respecting the activity thesis. So in escaping the rock of panpsychism, a computational theory of conscious experience inevitably runs straight into a hard place: the denial of the activity thesis.

Acknowledgments. Thanks to Tim Maudlin, Colin Klein, and an anonymous referee for very helpful comments.

Notes

[1] This assumes that the machinery that is called upon for the run on (I) is *physically distinct from* the machinery that would be called upon otherwise, so that the latter can be removed without removing the former. But while such a scenario might be a bit unusual in actual computing machines, it would be very odd and *ad hoc* for computationalists to say that their theory depends upon the absence of such a scenario.

[2] Thus Pylyshyn (1984: 55): "Very few of the physically discriminable properties of [a] machine are relevant to its computational function…. In fact, *any* variations in physical properties of distinct components, other than those few properties to which other designated components are meant to react in *particular* specified ways, can be said to be irrelevant to the machine's operation as a computer."

[3] Bishop (2002a, b) adds that since we can turn R_1 into R_2 by *gradually* deleting unused execution traces, qualia must fade over the series of systems between R_1 and R_2. Yet this, he implies, is the very result that Chalmers deplores in his own 'fading qualia' argument (Chalmers 1995). However, as Chalmers has countered (on his website), what he denies is that a system's qualia can fade *while its functional organization is held fixed*. R_1 and R_2 differ in their functional organization, so Chalmers is free to hold that the removal of the unactivated states *does* matter.

[4] Computationalists might deny this on the grounds that Klara has the wrong kind of functional architecture. They may say that experience requires not just that certain computations be performed, but that they be performed in a certain way on a certain sort of architecture. The brain's architecture is surely nothing like a Turing machine!

However, ploys of episodic instantiation are not restricted to Turing machines. (Maudlin used them in the interests of generality, but overlooked the limitations of their architecture.) All that needs to be shown, as I am about to explain, is that the counterfactuals supporting a wide range of a system's responses to a certain class of input can be falsified, while leaving untouched the machinery necessary for execution on one member of that input class. There is no reason to think that this sort of manipulation is only possible in non-sentient computing systems.

[5] Eric Barnes (1991) holds that Olympia lacks φ even when the Klara copies are *present*, because the input τ is not an 'active cause' of her activity, and hence she is not computing $\pi(\tau)$ at all. I have two remarks about this. Firstly, Barnes bases his argument on a dubious analogy between cognition and computation. While it is plausible that *cognizing an object* requires active causation by the object, that does not entail that *computing a function on an input* requires active causation by the input. And secondly, in any case, there are ploys of episodic instantiation in which the operations of the second system *are* actively caused by the input, as with Bishop's R_2.

[6] Valerie Hardcastle (1993) argues that while Olympia (with the Klara copies) may have experience φ, being a conscious *system* "requires more than simply exhibiting a subset of the possible phenomenal experiences" (¶25). Hardcastle's position is puzzling, however, for a computational theory will first and foremost be a theory of conscious *states*. A conscious system plausibly *just is* a system with conscious states, and Hardcastle effectively admits that Maudlin has identified a problem for such a theory. That is the kind of theory with which I am concerned.

[7] Similarly, both Chalmers (1996) and Copeland (1996) observe that since a wild instantiation is a singular event, picked out *post hoc*, its computation-mimicking activity is underwritten only by material rather than counterfactual conditionals – thus we may say that it is not *disposed* to undergo the same activity again under similar conditions.

[8] At one point he hints that episodic implementations that produce experience will require some dispositions to manifest *repeatedly*. This suggests the idea that a wild instantiation of such an implementation would be impossible because wild instantiations are too unstable to manifest the same disposition repeatedly. However, I do not think that Klein actually adopts this view. It is not a plausible view. Even if we accept that experiences entail computations in

which some dispositions manifest repeatedly, there is no reason to think that such computations *cannot* be mimicked by wild instantiations – only that such wild instantiations will be hugely unlikely.

[9] Another source of uncertainty is whether Klein thinks that wild instantiations possess the appropriate dispositions very fleetingly, or that they do not possess those dispositions at all. In the paper his emphasis on the importance of *standing* or *stable* dispositions implies the former, but in personal communication he asserts the latter.

[10] Indeed, in personal communication Klein says he prefers to say that they are *not* part of Olympia, and therefore that Olympia is not computing because the central unit on its own lacks a disposition that supports the relevant counterfactuals. (I will remark, in passing, that this seems not to fit with the episodic account of implementation. Surely Olympia, construed just as the central unit, *episodically implements* $\pi(\tau)$? After all, the point of the account is that it jettisons the need for all the counterfactuals associated with π to be supported.)

[11] Later in the footnote about Olympia, Klein seems to acknowledge the very distinction – between dispositions and activity – that I am claiming he wishes to elide. He says that if the Klara copies are part of Olympia, then since interfering with them interferes with Olympia's dispositions, there can be no change in her computational status "without a change in *either dispositional structure or actual activity*" (p. 150n, my emphasis). However, this cannot be right, for it contradicts his assertion in the main text that computational activity supervenes on actual activity. In personal communication Klein confirms the error: the footnote text should read '*both* dispositional structure *and* actual activity.'

[12] While my own notion of physical activity entails that it is multiply realizable, it is not functionalist or computationalist. Rather, I have in mind an intrinsic characterization of activity (cf. the intrinsic structural properties suggested by Pereboom 2002). For example, the physical activity of 'flexing' would be realizable by rubber hoses, metal paper clips, tree branches, and many other things – but it would nevertheless be characterized intrinsically, in reference to certain structural properties that all of these items possess at the time. Of course, I do not claim to know how certain kinds of physical activity would be able to produce conscious experience. But that can hardly be held against the idea, since no one knows how *any* particular non-mental property produces experience.

[13] Antony's argument is similar to Maudlin's and Bishop's, but targets functionalism rather than computationalism. As such, it faces more difficulty, for activity plays a less prominent role in functionalism than in computationalism. I believe that Antony's argument begs the question against the functionalist. In another paper ([unpublished MS, in progress), I attempt to revise his argument in a way that eliminates its dependence on the activity thesis.

[14] Someone might argue that *all* the neurons in the brain are actively involved in producing one's experience *all the time* because even activity that does not rise to the level of an action potential still contributes to experience in some way. Perhaps this is true. But even if it is, it doesn't help CTE. CTE claims that *no activity at all* is needed in

order for a neuron to play a part in supporting experience – for according to CTE, experience is (at least partly) a *relational* phenomenon. All it takes is for a neuron to stand in a certain abstractly-defined relation to some neurons that *are* active. So CTE can't look for help from the 'pan-activist' view just adumbrated.

[15] Thanks to Colin Klein for drawing my attention to the Wada test.

References

Antony, M. V. (1994). Against functionalist theories of consciousness. *Mind & Language*, *9*, 105-123.

Barnes, E. (1991). The causal history of computational activity: Maudlin and Olympia. *The Journal of Philosophy*, *88*, 304-316.

Bishop. M. (2002a). Dancing with pixies: strong artificial intelligence and panpsychism. In J. Preston & M. Bishop (Eds.), *Views into the Chinese room: new essays on Searle and artificial intelligence* (pp. 360-378). Oxford: Clarendon Press.

Bishop, M. (2002b). Counterfactuals cannot count: a rejoinder to David Chalmers. *Consciousness and Cognition*, *11*, 642-652.

Bishop, M. (2009). A cognitive computation fallacy? Cognition, computations and panpsychism. *Cognitive Computation*, *1*, 221-233.

Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Minnesota Studies in the Philosophy of Science, Vol. 9: Perception and Cognition* (pp. 261-325). Minneapolis: University of Minnesota Press.

Chalmers, D. J. (1995). Absent qualia, fading qualia, dancing qualia. In T. Metzinger (Ed.), *Conscious experience* (pp. 309-330). Schöningh: Imprint Academic.

Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese*, *108*, 309-333.

Chalmers, D. J. (n.d.). Responses to articles on my work: Mark Bishop. Retrieved June 11, 2008, from http://consc.net/responses.html#bishop.

Chrisley, R. L. (1994). Why everything doesn't realise every computation. *Minds and Machines*, *4*, 403-420.

Chrisley, R. L. (2006). Counterfactual computational vehicles of consciousness. Paper presented at 'Toward a Science of Consciousness 2006', Tucson, April 7th. [Powerpoint retrieved June 11, 2009, from http://e-asterisk.blogspot.com/2006/05/counterfactual-computational-vehicles.html.]

Copeland, B. J. (1996). What is computation? *Synthese*, *108*, 335-359.

Dretske, F. I. (1981). *Knowledge and the flow of information*. Cambridge, MA: M.I.T. Press.

Hardcastle, V. G. (1993). Conscious computations. *The Electronic Journal of Analytic Philosophy*, *1*. Retrieved July 23, 2008, from http://ejap.louisiana.edu/EJAP/1993.august/hardcastle.html.

Klein, C. (2008). Dispositional implementation solves the superfluous structure problem. *Synthese*, *165*, 141-153.

Lycan, W. L. (1987). *Consciousness*. Cambridge, MA: The MIT Press.

Maudlin, T. (1989). Computation and consciousness. *The Journal of Philosophy*, *86*, 407-432.

Pereboom, D. (2002). Robust nonreductive materialism. *The Journal of Philosophy*, *99*, 499-531.

Piccinini, G. (2008). Computers. *Pacific Philosophical Quarterly*, *89*, 32-73.

Putnam, H. (1988). *Representation and reality*. Cambridge, MA: The MIT Press.

Pylyshyn, Z. W. (1984). *Computation and cognition: toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417-424.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: The MIT Press.