DUBLIN CITY
UNIVERSITY
Ollscoil Chathair Bhaile Átha Cliath

Computing Machinery and Mentality*

Barry McMullin

December 1993

School of Electronic Engineering

TECHNICAL REPORT: bmcm9302

School Of Electronic Engineering,
Dublin City University,
Dublin 9,
IRELAND

Telephone: +353-1-704 5432
Fax: +353-1-704 5508
E-mail: McMullin@EENG.DCU.IE

*Invited paper presented at the workshop *Artificial Life: a Bridge towards a New Artificial Intelligence*, San Sebastian, December 10th and 11th, 1993. This is an abridged version of discussions first presented in Chapter 2 of my Ph.D. Thesis (McMullin 1992).

Abstract

I reconsider the status of *computationalism* (or, in a weak sense, *functionalism*): the claim that being a realisation of some (as yet unspecified) class of abstract *machine* is both necessary and sufficient for having genuine, full-blooded, *mentality*. This doctrine is now quite widely (though by no means universally) seen as discredited. My position is that, though it is undoubtedly an unsatisfactory (perhaps even repugnant) thesis, the arguments against it are still rather weak. In particular, I critically reassess John Searle's infamous *Chinese Room Argument*, and also some relevant aspects of Karl Popper's theory of the *Open Universe*. I conclude that the status of computationalism must still be regarded as *undecided*; and that it may still provide a satisfactory framework for research.

1 Introduction

> Many psychologists and brain scientists are embarrassed by the philosophical questions, and wish no one would ask them, but of course their students persist in asking them, because in the end these are the questions that motivate the enterprise.
>
> Dennett (1978, p. xiii)

In this paper I shall be concerned with the question of whether the research programme which goes under the title of *Artificial Intelligence*, or *AI*, is capable (even in principle) of solving any of the substantive problems posed by the existence of *minds.* My conclusion will be the weakest possible in the circumstances: I shall claim merely that the case against AI, or "computationalism" in the broadest sense, is *not (yet) proven.* It is quite enough for my purposes that the question still be open. Specifically, I do not propose to argue that AI demonstrably *can* solve any particular problem(s) of mentality. Or, if you wish, I accept that the case *for* AI (as an approach to mentality at least—I ignore any questions concerning technological *utility*) is, equally, not (yet) proven.

There is not, of course, enough space here to attempt a properly *comprehensive* review of these questions (nor, for that matter, would I be remotely qualified to attempt such a task). Instead I have made a somewhat iconoclastic selection of just two arguments to consider, those put forward by John Searle and by Karl Popper. I consider Searle's argument simply because it has become the touchstone for very much of the ongoing debate, and could hardly, therefore, be avoided. Popper's argument is less well known, and rather more difficult to assess. I take it up primarily because Popper is, arguably, the greatest philosopher of this century, and his discussions of these issues therefore demand the most serious and careful consideration.

2 Three Hypotheses

I shall state three related hypotheses, which will then serve as targets for criticism.

H_p **(Physicalism):** All mental states and events can, in principle, be completely reduced, without residue, to physical states and events. This is physicalism *simpliciter*, or so-called *token* physicalism—see, for example, Block (1980c, p. 296).

H_c **(Computationalism):** All mental states and events can, in principle, be completely reduced, without residue, to states and events of some universal computer. This is equivalent to at least some forms of *functionalism* (Block 1980a).

H_t **(Turing Test Computationalism):** The Turing Test (Turing 1950) can be passed by certain systems whose *putative* mental states and events can, in principle, be completely reduced, without residue, to states and events, of some universal computer. H_t is, essentially, a behaviouristic version of H_c.

H_c implies both H_p and H_t. H_c is the hypothesis of direct interest here; I have introduced H_p and H_t solely because any (alleged) refutations of these would also refute H_c.

3 Refuting Computationalism?

I now turn to two quite distinct attempted refutations of H_c. These are Searle's so called *Chinese Room* thought experiment, and the rather more general "dualist interactionist" argument for the causal openness of the physical world (which is to say, for the falsity of H_p, and thus, implicitly, of H_c also) presented by Popper & Eccles. It seems to me that these are substantial and challenging arguments, and I shall devote the following sections to considering them in some detail.

3.1 Searle's Chinese Room

> Searle's [Searle 1980] 'Chinese Room' argument against 'Strong AI' has had considerable influence on the cognitive science community ... it has challenged the computational view of mind and inspired in many respondents the conviction that they have come up with decisive, knock-down counterarguments ... Yet the challenge does not seem to want to go away ... Indeed, some have gone so far as to define the field of cognitive science as the ongoing mission of demonstrating Searle's argument to be wrong.
>
> Harnad (1989)

John Searle's original presentation of his Chinese Room argument was already accompanied by extensive peer commentary (Searle 1980). In the years

that have since passed, there has been a continuing stream of publication on the issue. A survey is provided by, for example, Harnad in the paper quoted above. Slightly more recently, *Scientific American* has hosted another instalment in the debate, with a restatement of his position by Searle, and an attempted rebuttal by P.M. Churchland and P. Smith Churchland (Searle 1990; Churchland & Churchland 1990).

In what follows, I shall take "Strong AI", as Searle terms it, as being equivalent to my H_c, and "Weak AI" as equivalent to my H_t.

Searle's contention is that H_c is false, and that this is demonstrable through a series of thought experiments. I shall describe only the simplest of these, and even that only very briefly.

Let there be a computer which (when suitably programmed) appears to instantiate the mentality of a Chinese speaking person (in something like the sense of the Turing Test). A person, ensconced in the so-called *Chinese Room*, could, given appropriate, purely formal, instructions, simulate the behaviour of this computer exactly. This Chinese Room would also, therefore, putatively instantiate the mentality of the Chinese speaking person. The "real" person carrying out the simulation is stipulated not to be a Chinese-speaker. If we now enquire of this person whether she understands any Chinese, she will say no. Therefore (?) there is no genuine Chinese mentality being realised by the Chinese Room, and therefore mentality cannot be reduced, without residue, to computational states and events. H_c has been refuted.

It is important to note that Searle *accepts* H_p, or, at least, something essentially equivalent to it:

> Can a machine have conscious thoughts in exactly the same sense that you or I have? If by "machine" one means a physical system capable of performing certain functions (and what else can one mean?), then humans are machines of a special biological kind, and humans can think, and so, of course machines can think. And, for all we know, it might be possible to produce a thinking machine out of different materials altogether—say, out of silicon chips or vacuum tubes. Maybe it will turn out to be impossible, but we certainly do not know that yet.
>
> Searle (1990, p. 20)

So, Searle's claim is that some sort of physicalist (H_p) theory is (or at least, may be) true—but that H_c is not that theory.

Searle is neutral with respect to H_t: indeed, the Chinese Room argument only works given the assumption that H_t may, in fact, be true (if H_t somehow actually proves to be false, then that automatically refutes H_c anyway, and the fact that the Chinese Room argument could no longer even be properly formulated would not matter—it becomes redundant with respect to the real problem, i.e. the truth or otherwise of H_c).

Now most, if not all, commentators on this issue can be divided into two groups:

- Those who hold that H_c is false, whether they agree with all of Searle's reasoning or not. Thus I include here, for example, Eccles (1980), who agrees with Searle's refutation of H_c, but disagrees strongly with Searle's uncritical acceptance of H_p (Eccles describes himself, following Popper, as a "dualist interactionist"—see Popper & Eccles 1977; I shall consider their views in more detail in section 3.2 below).

- Those who hold that H_c is true. Their basic position is that, since H_c is true, Searle *must* be wrong. They then go on, *in the light of this,* to try to identify precisely why Searle is, in fact, wrong. I consider that, if *any* of these particular commentators are right, it is those who advocate the so-called "systems reply". Briefly, this grants that the person in the Chinese Room *per se* does not have any Chinese understanding or mentality, but holds that the Room *as a systemic whole* (including the person inside) understands, or at least, might understand, Chinese—i.e. have "genuine" Chinese mentality. However, I shall not pursue the arguments for and against that position here.

My purpose in making this classification is to identify, by omission, a third possible position: that which holds that Searle's reasoning is wrong, and that, therefore, the status of H_c is simply *unaffected* by his argument: it remains a tentative hypothesis. This is the position I propose to adopt.

It is important to realise that this is a perfectly valid procedure, and is, if correct, preferable to a position of claiming that H_c is actually true. It is preferable in the basic sense that attempting to argue for the truth of the converse of a proposition is, in general, an *unnecessarily strong* way of attacking a supposed proof of the original proposition. But the procedure is doubly preferable in this particular case where any attempt to prove the truth of H_c inevitably undermines itself anyway—it is a variant of what Popper (1965, p. 217) has termed "the nightmare of the physical determinist". I suspect

that this may be at the root of Harnad's observation that, "Many refutations [of Searle's argument] have been attempted, but none seem convincing" (Harnad 1989, p. 5).

So, to reiterate, my claim is that Searle's reasoning is defective, and his conclusion (that H_c is false) is therefore *unwarranted*; but I do *not* suggest that H_c is, in fact true. My only claim is that its status is still open.

Briefly, the argument is this:

> H_c does not make the prediction which Searle ascribes to it (that the person in the Chinese room should, upon enquiry, report that she understands Chinese); in fact, H_c is *neutral* as to the outcome of the experiment. H_c cannot, therefore, be *refuted* by Searle's experiment—*no matter what its outcome!*

As far as I am aware, this argument is due to Drew McDermott, who introduced it in personal communication with Harnad; I have not identified any published version of precisely this idea. For myself I consider this argument to be concise, elegant and devastating. On the other hand, as Harnad stated in my opening quotation above, many have previously thought they had identified "decisive" arguments on this issue, but the debate rumbles on nonetheless (indeed, Harnad himself rejected this view of McDermott's, but I have been unable to understand his reasons).[1]

In any case, I now turn back to Searle's own arguments. Searle has, I think, been somewhat puzzled by the reception his ideas have had—at least in the AI community. He believes that his Chinese Room Argument is decisive against H_c, and yet there are many people who are unwilling to accept this. So he seeks an explanation of this. He finds a candidate explanation in the notion that some people may (mistakenly) think that H_t necessarily implies H_c. Therefore, anyone who accepts Turing's original argument for H_t (basically, a universal computer can realise any effective procedure—can "simulate" anything whose behaviour is sufficiently well specified—and there is no manifest *a priori* reason for supposing that human linguistic performance cannot be so specified) would interpret this as an argument for H_c also; and might therefore be convinced that Searle must be wrong in his refutation of H_c, even if they cannot identify exactly *why* he is wrong.

[1]Excerpts from this correspondence between Harnad and McDermott were distributed by Harnad through his electronic discussion group on the so-called *symbol grounding problem*; my discussion is based on a message dated `Sun, 13 May 90 23:11:40 EDT`.

Now even Searle himself is willing to accept the *possibility* that H_t may be true. So he perceives that part, at least, of his task should be to show how it might be that H_t could be true, and yet H_c could be false.

He does this by citing other phenomena (e.g. rainstorms) which can be perfectly well *simulated* by computers, but which plainly cannot be so *realised* (a simulated rainstorm cannot make you wet!). By analogy, he argues, there is no reason to suppose that the mere simulation of a mind (H_t) would actually cause a "real" mind to be called into existence (H_c)—(Searle 1980, p. 423).

My comment is simply to say that all this is certainly true, insofar as it goes, but it is not germane; at least, it is not germane to *my* disagreement with Searle.

Thus, I *do* say that, in a certain special sense, H_t *might* imply H_c; but this is not my reason for rejecting the Chinese Room Experiment, and it is not at all affected by spurious meteorological analogies (ironically, Searle himself warns against the dangers of wanton analogising—Searle 1990, p. 24). In fact, the situation is exactly opposite to that apparently envisaged by Searle.

I *start* with a rejection of the Chinese Room argument (following McDermott, as explained above). I therefore also, implicitly, reject Searle's alleged distinction between mere mind-like behaviour (H_t) and real minds (H_c). I then conjecture that, in the absence of some alternative criterion for distinguishing H_c from H_t (i.e. independently of the Chinese Room Experiment) the two are (*pro tem*) identical (i.e. the Turing Test is a *bona fide* test for mentality); and in this very special, degenerate, sense, it can actually be technically correct, although not very illuminating, to say that H_t implies H_c (rainstorms notwithstanding).

Or to put it another way, Searle's analogy only begins to make sense if we already accept that minds are entities like rainstorms, whose realisation demands certain specific, physical, causal powers, and are *not* entities like computers (or, if you prefer, computations) which can be realised by more or less arbitrary physical systems; but if we already accepted *that,* we would have already accepted the falsity of H_c, and the analogy would be unnecessary. It seems that, whichever way you look at it, Searle's discussion of simulation versus realisation does not add anything to the original argument.

Of course, on this scenario, I should stress that I take H_t (and therefore, still, H_c) to be strictly conjectural and unproven.

Finally, in concluding this discussion of the Chinese Room argument, I should emphasise my admiration for the boldness of Searle's idea—that it might be possible to refute H_c *prior* to coming to any conclusion on H_t. Unfortunately, Searle's particular idea for doing this does not work.

3.2 Dualist Interactionism

It seems to me that, almost by definition, the only (realist) alternative to physicalism is some kind of pluralism; that is, one must suppose that there exist distinct classes of entity which interact with each other (they are *real*) but which are not reducible to the class of physical entities (supposing, for the sake of the argument, that the latter class could be well defined in an unproblematic way—*cf.* Block 1980c, p. 296). As far as mentality is concerned, this means a *dualist interactionist* position: holding that mental events are genuine entities, having causal effects on physical entities, but not themselves reducible to physical entities.

There is a distinction to be noted here between merely holding that physicalism is unproven (or even "unlikely"), and holding that it is actually false—i.e. *positively* advocating a dualist position.

Such a dualist position seems, however, not to be currently fashionable in the philosophy of mind. The *only* substantive contemporary example cited by Hofstadter and Dennett, in their extensive annotated bibliography of the field (Hofstadter & Dennett 1981, pp. 465–482), is that of Popper & Eccles (1977); I shall therefore give careful attention to a consideration of their position.

3.2.1 Eccles Neurophysiological Perspective

Eccles professes himself a dualist interactionist, but, as far as I have been able to establish, does not marshal any particular arguments in favour of this position. In his joint book with Popper, this issue is primarily dealt with in Chapter E7, where he expressly describes his purpose, not as the establishment of dualism as such, but as "the development of a new theory relating to *the manner* in which the self-conscious mind and the brain interact" (Popper & Eccles 1977, p. 355, emphasis added). That is, Eccles adopts the dualist interactionist hypothesis, *for whatever reasons,* and goes on to explore some of the consequences of this hypothesis; specifically, enquiring into the *nature* of the interaction between mind and brain.

I shall presume, though Eccles appears not to state it explicitly, that he relies on Popper for the prior establishment of the dualist position: his own rôle is then to consider some more specific implications of this general position. My task thus reduces to that of considering Popper's arguments alone; to the extent that I claim they are flawed, the considerations raised by Eccles are at least premature, if not irrelevant.[2]

3.2.2 Popper on AI

Popper is, at least, unambiguous in his view of what I have called H_c—he holds that it is false:

> I have said nothing so far about a question which has been debated quite a lot: whether we shall one day build a machine that can think. It has been much discussed under the title "Can Computers Think?". I would say without hesitation that they cannot, in spite of my unbounded respect for A.M. Turing who thought the opposite ... I predict that we shall not be able to build electronic computers with conscious subjective experience.
>
> Popper & Eccles
> (1977, Chapter P5, pp. 207–208)

Popper is less clear cut on H_t:

> Turing [Turing 1950] said something like this: specify the way in which you believe that a man is superior to a computer and I shall build a computer which refutes your belief. Turing's challenge should not be taken up; for any sufficiently precise specification could be used in principle to programme a computer. Also, the challenge was about behaviour—admittedly including verbal behaviour—rather than about subjective experience.
>
> Popper & Eccles
> (1977, Chapter P5, pp. 208)

It seems that Popper accepts Turing's argument as showing that a suitably programmed computer may well be able to exhibit behaviour sufficient to pass the Turing Test (say); but considers *therefore* that there is little point in pursuing this. In particular, it will not necessarily endow a computer with "conscious subjective experience".

[2] Eccles does make one other point that might be taken as a rationale for his dualist position—that he is "a believer in God and the supernatural" (Popper & Eccles 1977, p. VIII); but he does not expand any further on this, and thus there is no basis for substantive discussion here.

Thus far, Popper's position is quite comparable to that of Searle. However, his arguments for this position are entirely different, as we shall see.

3.2.3 The Open Universe

Popper explicitly rejects physicalism, in all its manifestations, including what I have termed H_p. This is quite different from Searle who, as we saw, seems willing to accept the general idea of physicalism, rejecting only the special case represented by H_c.

Popper describes himself as a "dualist interactionist" with respect to the mind-body problem. However, he presents this in the context of his more general philosophy of the *Open Universe,* or what we might term a "pluralist" (rather than merely dualist) cosmology. That is, Popper holds that there exist, in the real universe, a variety of distinct classes of entities which are mutually interacting, but which are not reducible to each other; and that, furthermore, new irreducible classes of entity can, and do, *emerge* over time.

In particular, Popper has identified three specific classes of entities which, he claims, are not reducible to each other, and which he terms *Worlds.*

World 1 is the conventional world of unproblematic (?) physical entities. *World 2* is the world of subjective mental entities such as emotions, intentions, sensations, ideas, thoughts etc. Finally, *World 3* is the world of:

> ... products of the human mind, such as stories, explanatory myths, tools, scientific theories (whether true or false), scientific problems, social institutions, and works of art.
>
> Popper & Eccles
> (1977, Chapter P2, p. 38)

Thus, Popper specifically claims that World 1 and World 2 interact (they both contain *real* entities in good standing), but that they are mutually irreducible. This establishes his *dualist* position on the mind-body problem.

Popper has described the general idea of the Open Universe, and the Worlds 1, 2 and 3, in a wide variety of his writings. However, in what follows I shall restrict myself, for the most part, to the presentation of Popper & Eccles (1977), as this is where Popper explicitly relates this idea to the problem of artificial intelligence (or, at least, of artificial mentality).

Popper's attack on physicalism is two pronged: on the one hand, he identifies specific difficulties with a purely physicalist position; and on the other, he argues positively in favour of the dualist position. My criticism will therefore be similarly twofold.

3.2.4 Arguing Against Physicalism

Firstly, let me consider the specific difficulties alleged for physicalism. Popper provides a survey of varieties of physicalism, and adduces slightly different arguments against them. For my purposes, it is sufficient to concentrate on one specific variant, the (token) *identity theory* (Popper & Eccles 1977, Chapter P3, Sections 22–23). Popper considers this the most difficult version of physicalism to rebut, going as far as to grant that, viewed in isolation, it *may* be true. However he claims that it is incompatible with Darwinism, and then argues that, since we must therefore choose between these two theories, we should prefer to retain Darwinism rather than physicalism.

My position is that Popper is mistaken in claiming that the identity theory (which is essentially equivalent to my H_p) is incompatible with Darwinism. Popper himself admits that his argument here is less than intuitively clear. It will require some care to deal properly with it—both to do justice to it in the first place, and then to answer it convincingly.

Popper's argument is that, under the identity theory, Darwinism is powerless to explain the *evolution* of mental entities, *per se.* This is so because:

- A Darwinian explanation can only work if the evolved entity has physical effects (roughly, it must positively affect the reproductive success of the carrier organisms).
- In the final analysis, under the identity theory, the mental entity can be shown to have physical effects *only* by replacing it with the (putative) physical entities with which it is identical.
- Such a purely physical Darwinian explanation, which has been shorn of all mental entities may, indeed, be valid. It will then properly explain why certain purely physical entities can evolve (i.e. because they are favoured by natural selection).
- However, since this explanatory scheme no longer contains any mental entities it is powerless to shed any light on why the (physical) entities which evolve are, in fact, identical with some mental entity.
- To put it another way, we would have a Darwinian explanation for the evolution of certain physical entities; we would *separately* know that

these are identical to some mental entity; but this latter fact would have played no rôle in the evolutionary explanation. Thus, we could not then claim that the physical entities in question had evolved *by virtue* of this identity, nor of any properties of the mental entity, as such. We would have an explanation for the evolution of certain physical entities, but the fact that these are *also* correlated with (are identical to) some mental entities would stand as an independent, unexplained, and inexplicable, phenomenon. Indeed, according to our explanation they would have evolved in just the same way, even if they were *not* identical with some mental entity.

- That is, a Darwinian explanation for the specifically mental character of certain evolved physical entities is impossible. We would require some alternative explanatory principle, *in addition to Darwinism*, to address this.

- The incompatibility between the identity theory and Darwinism resides precisely in this result: that Darwinism would not be effective in explaining the evolution of mental entities.

I believe I have here stated Popper's argument in about as strong and as clear a form as is possible. I should add that Popper (Popper & Eccles 1977, Chapter P3, p. 88) also refers to a similar argument having been independently formulated by Beloff (1965).

I claim that the flaw in the argument is simply this: it goes through if and only if the characteristics of the physical entities which are relevant to their Darwinian selection are independent of (uncorrelated with) the characteristics which are relevant to their identification with some mental entity. To put it another way, an identification between a mental entity and some physical entities will, in the last analysis, require the physical entities to have some specific physical characteristics—otherwise the identification would be unwarranted. These physical characteristics may not be sufficient for the particular identification, but they would be necessary. Once this much is granted, it is unproblematic to incorporate these particular physical characteristics, which are essential elements of the identification, as factors in a Darwinian explanation of the evolution of the (identified) mental entity.

To be specific, suppose that we have available to us a conjectural reduction of the entire mentality of some person to "unproblematic" physical entities: that is, we have a procedure for making identifications between the person's mental states and events and some physical states and events. A *necessary* (though not sufficient) condition for accepting this reduction, or system of identifications, is that the physical effects that result must be more or less consistent with the identified mental states and events—for example, the physical linguistic behaviour implied by the purely physical model must be consistent with the supposed mental states which correspond to it. To modify slightly an original example due to Fodor (1976, p. 199), one might postulate some particular identification which then turns out to have the property that a mental state of *believing that it will rain* predicts the consequent occurrence of the physical utterance "there aren't any aardvarks any more"; but one would then conclude that this identification between beliefs and physical states is, to say the least, suspect!

Ultimately, the core of Popper's argument seems to be this: if World 1 is causally closed (H_p is true), then Darwinism can, at best, provide an explanation of the evolution of certain physical phenomena, but these, in themselves, will have no *necessary* connection with subjective mental experience. Indeed, it seems to be apparent from Popper's criticism, already quoted, of the notion of Turing Testing, that he envisages that a system *could* well exhibit extremely complex behaviours, up to and including human level linguistic behaviours, and yet completely lack mentality; in a phrase commonly invoked by Harnad, it may be the case that, despite all appearances to the contrary, there could simply be "nobody home". If this is indeed possible—if the physical (including linguistic) manifestations of mentality can be had in the absence of mentality proper—then mentality would, from a Darwinian point of view, be redundant, and Darwinism would be incapable of explaining its evolution. But, if this *is* Popper's point, it seems to beg the question at issue: the idea of H_p (and, more specifically, of H_c) is precisely to conjecture that mentality proper—in the sense of "conscious subjective experience"—*is* an inevitable correlate of certain physical behaviours. Now this conjecture may surely be mistaken, but it can hardly be criticised by an argument which already assumes it to be false.

It seems to me that the essence of the problem here for Popper, as previously for Searle, is to find an effective wedge to drive between H_c and H_t—for they both wish to accept the latter (tentatively, at least) but still reject the former. But once seen in this light, we can recognise that it is a very tall order indeed: it requires, more or less, a solution to the "other minds" problem—a basis for discriminating the mere "appearance" of mentality from "genuine"

mentality. While Popper's approach is very different from Searle's, I cannot see that he is ultimately any more successful.

3.2.5 Arguing For Dualism

Next let us consider Popper's *positive* argument in favour of dualist interactionism (Popper 1973; Popper & Eccles 1977, Chapter P2).

The core of the argument is the claim that there exist at least some World 3 entities which are real (i.e. which interact, albeit indirectly, with World 1) but such that they are demonstrably *not* reducible to physical entities, i.e. are not *identifiable* with World 1 entities (they are "unembodied" in Popper's terms).

This would be enough to establish that the strictly physicalist view must be false. It would not, in itself, establish *mind-body* dualism, as such, i.e. the irreducibility of *World 2* to World 1. Popper completes the argument by pointing out that, in general, World 3 interacts with World 1 only through the mediation of World 2; therefore (so the argument goes), since World 3 itself is irreducible to World 1, and World 2 can interact with World 3, a capacity not exhibited by World 1 in general, then World 2 must *also* be irreducible to World 1.

I suggest that this latter argument is, in fact, defective. To see this, note that, under the identity theory (which Popper accepts "may" be true), the distinction between the mental and the physical is simply that certain states or organisations of World 1 entities do exhibit precisely the characteristics of World 2 entities, and, in this way, World 2 may be reduced to World 1. To apply this theory in Popper's scheme, we would simply stipulate that these distinguishing ("mental") characteristics of certain World 1 entities must include the ability to "grasp", as Popper puts it, World 3 entities. Popper has not offered any detailed theory of this interaction, which might show that it is beyond the ability of *some* such World 1 entities. Therefore, Popper has failed to justify the claim that interaction cannot happen *directly* between (unembodied) World 3 entities and (any) World 1 entities, and so has failed to establish the irreducibility of World 2 to World 1, as required for mind-body dualism.

The flaw in Popper's argument is, then, that he (implicitly) proceeds from the premise that *certain* unembodied World 3 entities cannot interact directly with *certain* World 1 entities to the conclusion that unembodied World 3 entities cannot interact directly with *any* World 1 entities (such as minds, or rather, under the identity theory, the putative World 1 entities which are identifiable with minds). In taking this step he *assumes* the irreducibility of World 2 (i.e. the non-existence of World 1 entities which are identifiable with minds), which is precisely what he is purporting to establish.

However, this outcome is actually still peculiarly unsatisfying. We see that Popper's conclusion of mind-body dualism is unwarranted, because one particular step in his argument is defective. Perhaps this is enough: I claim to have provided a sufficient basis to refute Popper's argument for mind-body dualism, which is all I really sought to do. But: it involves attacking Popper on the *weakest* element of his argument, while still leaving his central, substantive, point unchallenged.

This central point is the claim that World 1 is causally *open*—that there exist entities which are demonstrably not reducible to World 1 entities, but which are perfectly real in the sense of *altering* the behaviours of some World 1 entities from what would be predicted based solely on their interactions with the rest of World 1.

It would be much more satisfactory if one could sustain a challenge against Popper's argument for an Open Universe as such, rather than relying on a rather technical nicety in how he has applied it to the issue of mind-body dualism. This is precisely what I shall now try to do.

The critical step is Popper's claim that certain World 3 entities are "unembodied", i.e. irreducible to World 1 (or World 2, for that matter), but, nonetheless, have definite causal effects on World 1 (via World 2).

The first part of this is unobjectionable: Popper is the originator of the World 3 concept, so he is surely entitled to include within it whatever he wishes. In particular, he may include things like *unproved theorems*: that is, statements which are, in fact, true (relative to some system of axioms) but for which no one has yet actually found a proof. By definition, such things are, indeed, unembodied—there do not exist any World 1 or World 2 entities correlated with them.

It is the second part of Popper's claim that seems to me to be potentially problematic: the assertion that such unembodied World 3 entities are *real*, in the sense of interacting directly with World 2, and thus indirectly (at least) with World 1. Popper deals explicitly with this issue as follows:

> ... Thus a not yet discovered and not yet embodied logical problem situation may prove decisive for our thought processes, and may lead to actions with repercussions in the physical World 1, for example to a publication. (An example would be

> the search for, and the discovery of, a suspected new proof of a mathematical theorem.)
>
> Popper & Eccles
> (1977, Chapter P2, p. 46)

If I understand him correctly, Popper's point here is that the truth of a mathematical theorem (for example) is an objective World 3 fact which is independent of any embodiment in World 2; it is, indeed, as objective as any World 1 fact. In particular, it is intersubjectively testable. Such tests are always fallible of course—but so too are tests of supposed World 1 "facts". Since these World 3 facts can exist and persist despite not being embodied, they evidently (?) cannot be reduced, without residue, to World 2 or World 1 entities; but since they *can* interact with World 2 (or be "grasped"), and thus with World 1, they are surely *real*. Popper's conclusion is then that World 1 cannot be causally closed.

This is a highly original and bold argument. It is, intuitively, quite compelling. And yet, when I examine it critically, it seems to me that it has very little substance, and cannot possibly be made to bear the burden which Popper attempts to place upon it.

Let us consider Popper's own favoured example: the truth of a mathematical theorem. This objective World 3 entity may be said to "interact" with a mathematician in the sense of constraining her results; she will not, in particular, be able to prove the theorem, nor any of its corollaries, *false*, no matter how hard she may try; the reality of the theorem may be said to manifest itself through the failure of such attempts. This is so, regardless of whether the mathematician ever explicitly conjectures, even, that this theorem exists. Let me stipulate, then, that this establishes the "reality" of the theorem.

The *irreducibility* of the theorem is separately held to follow from the fact that, at a given time, there may be nobody at all (no World 2 entities) who have yet even conjectured that it may hold, so there are not even any *candidate* World 2 entities as targets for a reduction (and thus, surely, there are no World 1 candidates either). But this claim is just wrong.

The theorem, if it *is* a theorem, is already implicit in the axioms of the system under study; it may be said to exist at all (in Popper's sense) only when some such axioms have been already *adopted*. That being the case, there is a perfectly good sense in which the theorem may be "reduced" to the *axioms*; and (by hypothesis) the axioms *are* already embodied, and thus *are* potentially reducible to World 2 (and ultimately even World 1) entities.

The point can be made more definite by replacing Popper's mathematician by a theorem proving *machine*. Such machines have indeed been built. By Popper's own hypothesis, such machines lack mentality, so they are not World 2 objects. Yet they can interact with, be constrained by, or even "grasp", the truth of a theorem in precisely the sense outlined above for a (human) mathematician. And they do so simply because this World 3 object, this truth of a theorem, is no more and no less than a product of the inference rules with which the machine was originally equipped. But the system in question here is a paradigm example of a causally closed physical (World 1) system. While it is true that, initially, the machine has no explicit embodiment of the theorem (even as a conjecture), this plainly does *not* establish (*pace* Popper) that the theorem is irreducible to World 1, or that the machine, *qua* World 1 entity, must be causally open to some influences which are not in World 1.

However, I am not sure that this quite exhausts Popper's argument yet. Popper is well aware of the possibility of theorem proving machines (though I am not aware of his having analysed their implications in just the way I have suggested above). Thus, even before he had fully formulated the concept of World 3, he made the following remark (this originally dates from c. 1957):

> A calculator may be able to turn out mathematical theorems. It may distinguish proofs from non-proofs—and thereby certain theorems from non-theorems. But it will not distinguish difficult and ingenious proofs and interesting theorems from dull and uninteresting ones. It will thus 'know' too much—far too much—that is without any interest. The knowledge of a calculator, however systematic, is like a sea of truisms in which a few particles of gold—of valuable information—may be suspended. (Catching these particles may be as difficult, and more boring, than trying to get them without a calculator.) It is only man, with his problems, who can lend significance to the calculators' senseless power of producing truths.
>
> Popper (1988, pp. 107–108)

This suggests to me a different, and more nebulous, interpretation of Popper's ideas. While I believe that the existence of theorem proving machines (even those proving uninteresting theorems!) adequately rebuts Popper's later, specific, claim that

"unembodied" theorems are necessarily irreducible to World 2 or World 1, it seems that Popper might not wish to rely on that argument anyway—that he has a much more general notion of an irreducible World 3 in mind. This is borne out, to an extent, in the following comment:

> There is no doubt in my mind that the worlds 2 and 3 do interact. If we try to grasp or understand a theory, or to remember a symphony, then our minds are causally influenced; not merely by a brain-stored memory of noises, but at least in part by the autonomous inner structures of the world 3 objects which we try to grasp.
>
> Popper (1973, p. 25)

To return again to the mathematician, it seems that Popper may wish to claim something much stronger than anything I have so far discussed. He may conceivably mean something like the following: that the objective existence of a theorem may change the pattern of the mathematician's thoughts so that (for example) she moves towards its formulation (or proof), in a way that is *not* already implied by her prior thoughts–i.e. in a way above and beyond the explanatory power of purely World 2 entities (noting of course, that the relevant World 2 entities will presumably be embodying certain World 3 entities). This should be contrasted sharply with a claim merely that the mathematician's *suspicions* or *intuitions* about the theorem affected her thought processes (as they undoubtedly would); for suspicions and intuitions are common or garden World 2 objects (presumably correlated with World 3 entities—but, by definition then, *these* are *already* embodied).

But it should be clear that any such interaction between unembodied World 3 entities and World 2 must be, at best, conjectural—one possible interpretation of the example of the mathematician, but not at all a conclusion from it. Indeed, if we apply Popper's own criteria for the evaluation of scientific theories, we should say that the hypothesis that unembodied World 3 entities do *not* have such causal effects on World 2 has a greater *content* (and thus corroborability) than its converse, and, in the absence of some evidence that it has actually been refuted (and none is offered, that I can see) should be preferred, even if only for the time being.

But the ramifications run deeper: such interactions between World 3 and World 2 would be completely inconsistent with the rest of Popper's evolutionary epistemology. They would be tantamount to a form of Lamarckian instruction by World 3 of World 2—i.e. Lamarckism applied to the evolutionary growth of an individual's subjective knowledge. This is something that has been resolutely opposed by Popper in the case of knowledge of World 1 (he has dubbed it the "bucket" theory of knowledge—Popper 1949; 1970), and I see no reason why his arguments should have any less force in the case of our knowledge of World 3. I therefore conclude that this cannot, after all, be a plausible interpretation of Popper's position.

It is important to note that none of my discussion here attempts to deny the *reality* of World 3 (an attack anticipated by Popper). I claim only that Popper has not established the *irreducibility* of World 3 to World 2 (and thus, possibly even to World 1). World 3 is still a perfectly meaningful and useful idea; as long as we admit that its reducibility is an open question, and that the hypothesis that it *is* reducible is actually stronger (has greater content) than the converse, and is currently a preferable basis for research.

4 Conclusion

In summary, my claim in this paper is that computationalism has not (yet) been definitively refuted; in particular, two distinct kinds of argument, by Searle and Popper respectively, purporting to achieve such a refutation, are flawed.

I should add a further comment, though I do not have space here to elaborate upon it: I consider that physicalism in general, and computationalism in particular, are irredeemably repugnant to human values and to the dignity of mankind. It seems to me that this kind of view precisely underlies the *intuitive* conviction of those, like Popper and Searle, who hold that H_c is definitely false. It should be clear that I completely share this intuitive conviction; I will confess, if that is the correct word, to being a *metaphysical* dualist.

However: the point at issue is how we might proceed *beyond* intuition. This raises what is almost a refrain of Popper himself:

> I regard intuition and imagination as immensely important: we need them to invent a theory. But intuition, just because it may persuade and convince us of the truth of what we have intuited, may badly mislead us: it is an invaluable helper, but also a dangerous helper, for it tends to make us uncritical. We must always meet

> it with respect, with gratitude, and with an effort to be severely critical of it.
>
> Popper
> (1988, Preface 1982, p. xxii)

Both Popper and Searle have attempted to proceed by supporting their intuitions with definite arguments—arguments which come close to having a scientific rather than a metaphysical character. If these arguments were acceptable—if H_c, in particular, were thereby refuted—then further investigations within the computationalist framework (such as, for example, attempts to *realise* Turing Test capability with computational systems) could only have technological significance; such investigations, though potentially valuable in their own right, would no longer directly bear on what I have elsewhere (McMullin 1992, p. 5) called *Popper's Problem*—the cosmological problem of understanding the world and our place in it. Thus, if one wished to remain focused on this latter problem then one would be led, instead, to proceed with a programme of research which reflected and incorporated the refutation of computationalism. Such an approach might be typified by the work of Eccles on the "liaison" between mind and brain, for example.

But I have claimed that the arguments put forward by Searle and Popper are flawed, and do not support the conclusions claimed. In particular, while I remain intuitively convinced of the falsity of H_c, this remains, for me, a *merely* intuitive belief. So the question remains of how best to proceed. Somewhat ironically, I think Popper has already suggested at least one possible answer to this:

> ... as a philosopher who looks at this world of ours, with us in it, I indeed dispair of any ultimate reduction. But as a methodologist this does not lead me to an anti-reductionist research programme. It only leads to the prediction that with the growth of our attempted reductions, our knowledge, and our universe of unsolved problems, will expand.
>
> Popper (1974, p. 277)

The programme of computationalism—of attempting to realise or synthesise the "appearances" (at least) of mentality by computational means—is an essentially reductionist one. Like Popper, I too do not expect any kind of ultimate success from this effort. But our failures, and the precise mechanisms of these failures, may be extremely interesting, and perhaps even revealing. There is thus every reason to pursue this programme of "methodological computationalism", despite our pessimism about its potential for "success"—just so long as we can avoid dogmatism, and continue to be critical of it.

References

Beloff, J. 1965. The Identity Hypothesis: A Critique. *In:* Smythies, J. R. (ed), *Brain and Mind*. London: Routledge & Kegan Paul. As cited by Popper & Eccles (1977).

Block, Ned. 1980a. Introduction: What Is Functionalism? *Pages 171–184 of:* (Block 1980b).

Block, Ned (ed). 1980b. *Readings in the Philosophy of Psychology.* Vol. 1. London: Methuen and Co.

Block, Ned. 1980c. Troubles with Functionalism. *Chap. 22, pages 268–305 of:* (Block 1980b). Reprinted, with revisons, from C. W. Savage, ed., *Perception and Cognition. Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science,* vol. 9, pp. 261–325, Minneapolis: University of Minnesota Press, 1978.

Boden, Margaret A. (ed). 1990. *The Philosophy of Artificial Intelligence.* Oxford Readings in Philosophy. Oxford: Oxford University Press.

Churchland, Paul M., & Churchland, Patricia Smith. 1990. Could a Machine Think? *Scientific American*, **262**(1), 26–31.

Dennett, Daniel C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology.* Brighton, Sussex: The Harvester Press Limited. Harvester Press edition first published in 1981.

Eccles, John C. 1980. A Dualist-Interactionist Perspective. *The Behavioral and Brain Sciences*, **3**, 430–431. Commentary on (Searle 1980).

Fodor, Jerry A. 1976. *The Language of Thought.* Hassocks, Sussex: The Harvester Press Limited. First published in the United States of America by Thomas Y. Crowell Company, Inc.

Harnad, Stevan. 1989. Minds, Machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, **1**, 5–25.

Haugeland, John (ed). 1981. *Mind Design.* Cambridge: MIT Press.

Hofstadter, Douglas R., & Dennett, Daniel C. (eds). 1981. *The Mind's I: Fantasies and Reflections on Self and Soul.* Harmondsworth, Middlesex: Penguin Books Ltd. Published in Penguin Books 1982.

McMullin, Finbarr (Barry) Vincent. 1992. *Artificial Knowledge: An Evolutionary Approach.* Ph.D. thesis, Ollscoil na hÉireann, The National University of Ireland, University College Dublin, Department of Computer Science.

Popper, Karl R. 1949. The Bucket and the Searchlight: Two Theories of Knowledge. *Pages 153–190 (Appendix 1) of:* (Popper 1979). First published 1949 (in German) as "Naturgesetze und theoretische Systeme" in Moser, Simon (ed), *Gesetz und Wirklichkeit.*

Popper, Karl R. 1965. Of Clouds and Clocks. *Chap. 6, pages 206–255 of:* (Popper 1979). This was the second Arthur Holly Compton Memorial Lecture, presented at Washington University on 21 Apr. 1965.

Popper, Karl R. 1970. Two Faces of Common Sense: An Argument for Commonsense Realism and Against the Commonsense Theory of Knowledge. *Chap. 2, pages 32–105 of:* (Popper 1979). This is a revised and expanded version of a talk first given by Popper in 1970, to his former Seminar.

Popper, Karl R. 1973. Indeterminism is not Enough. *Encounter*, **40**(4), 20–26. A revised version of this essay appears as *Addendum 1, pages 113–130, of:* Popper (1988).

Popper, Karl R. 1974. Scientific Reduction and the Essential Incompleteness of All Science. *Chap. 16, pages 259–284 of:* Ayala, Francisco Jose, & Dobzhansky, Theodosius (eds), *Studies in the Philosophy of Biology.* London: The Macmillan Press Ltd. A revised version of this essay appears as *Addendum 2, pages 131–162, of:* Popper (1988).

Popper, Karl R. 1979. *Objective Knowledge: An Evolutionary Approach.* Oxford: Oxford University Press. Revised edition (reprinted with corrections and a new appendix 2). First edition published 1972.

Popper, Karl R. 1988. *The Open Universe: An Argument for Indeterminism.* London: Hutchinson. From the *Postscript to the Logic of Scientific Discovery*, edited by W.W. Bartley, III. First edition published 1982. Note that the original text dates largely from the period 1951–1956.

Popper, Karl R., & Eccles, John C. 1977. *The Self and its Brain: An Argument for Interactionism.* London: Routledge & Kegan Paul plc. First published 1977, Berlin: Springer-Verlag. This edition first published 1983.

Searle, John R. 1980. Minds, Brains, and Programs. *The Behavioral and Brain Sciences*, **3**, 417–457. Includes peer commentaries. Also reprinted (without commentaries) as *Chap. 10, pages 282–305 of:* (Haugeland 1981) and as *Chap. 3, pages 67–88 of:* (Boden 1990).

Searle, John R. 1990. Is the Brain's Mind a Computer Program? *Scientific American*, **262**(1), 20–25.

Turing, Alan M. 1950. Computing Machinery and Intelligence. *Mind*, **LIX**(236), 433–460. Also reprinted as *Chap. 2, pages 40–66 of:* (Boden 1990).