Chapter 2

Artificial Mentality

2.1 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 (1978b, Introduction, p. xiii)

> In coming to grips with the idea of a natural system, we must necessarily touch on some basic philosophical questions ... This is unavoidable in any case, and must be confronted squarely at the outset of a work like the present one, because one's tacit presuppositions in these areas determine the character of one's science.
>
> Rosen (1985a, p. 45)

This chapter is concerned with the philosophical milieu in which the rest of the Thesis will be unfolded. More particularly, it is 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.*

This is no idle concern. As we shall see, a variety of critics, most notably Searle and Popper, have suggested that the answer to the question is a more or less simple *No*—that AI *cannot* illuminate the problems of mentality. If they were correct in this assessment it would represent a limitation, at the very least, on the applicability of subsequent discussions in the rest of the Thesis. This is so because, although I eschew many of the conventional tools and techniques associated with AI research, the work I describe still falls within the essentially *computational* paradigm which identifies AI as a field. It is as well to confront this issue at the outset.

My objective then, is to confound at least some of the critics of AI.

Having said that, let me immediately emphasise that my conclusion will be the weakest possible in the circumstances: I 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.

I suspect that this agnostic position is implicitly shared by most workers in AI; however, as Rosen points out in the quotation above, it is best to be explicit about such preconceptions.

2.2 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.

H_c **(Computationalism):** All mental states and events can, in principle, be completely reduced, without residue, to computational states and events, of some universal computer.

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 computational states and events, of some universal computer. H_t is, essentially, a behaviouristic version of H_c.[1]

[1] I shall review the Turing Test in detail in the next chapter. For present purposes, the following formulation is adequate: a system passes the Turing Test if, based on purely linguistic interrogation (e.g. via teletype), but spanning arbitrary topics, a competent judge mistakes it for a person.

H_c implies both H_p and H_t. Thus the following scenarios are logically conceivable:

- A refutation of H_c would be neutral with respect to both H_p and H_t.
- A refutation of H_p would be neutral with respect to H_t, but would constitute a *de facto* refutation of H_c.
- Similarly, a refutation of H_t would be neutral with respect to H_p, but would constitute a *de facto* refutation of H_c.

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

There are, of course, many other relevant hypotheses closely related to those I have introduced here, but with varying flavours and technicalities. However, in general, I deliberately overlook such finer distinctions in what follows, because they seem to be unnecessary refinements for the relatively modest purposes I have in mind.

2.3 A Personal Bias

> ...Yet machines are clearly not ends in themselves, however complicated they may be. They may be valuable because of their usefulness, or because of their rarity; and a certain specimen may be valuable because of its historical uniqueness. But machines become valueless if they do not have a rarity value: if there are too many of a kind we are prepared to pay to have them removed. On the other hand, we value human lives in spite of the problem of overpopulation, the gravest of all social problems of our time. We respect even the life of a murderer.
>
> Popper & Eccles (1977, Chapter P1, p. 4)

Before proceeding to consider criticisms of H_p and H_c, I should like to declare an element of personal bias: I side with those who hold that physicalism, whether in the plain form of H_p, or the more specific form of H_c, is utterly and irredeemably repugnant to human values. I shall therefore digress briefly to document just why I continue to regard physicalism with such distaste.

2.3.1 *Why* Physicalism is (Still) Repugnant

> *Physicalism is repugnant because it denies the freedom and responsibility of man.*

This is hardly a novel or original view, though it may have become less fashionable to speak of it (in the context of AI, at least). Indeed, some will, no doubt, consider me naïve to persist in it. However, I believe that this view, though hackneyed, is essentially correct. It has been effectively argued as such by, for example, Popper (Popper 1965; 1988; Popper & Eccles 1977).

Briefly, the physicalist hypothesis may be viewed as equivalent to the claim that the physical world is causally *closed* (which is, of course, not at all the same thing as claiming that the physical world is *deterministic*). This being so, mental states and events (i.e. minds, as such) can, in principle, be *dispensed* with in any description or analysis of physical states and events.

Minds may, of course, still be convenient devices for summarising certain (physical) phenomena. That is, minds may usefully be deployed in describing certain "law-like" physical behaviours. Indeed, it seems to me that it could only be by virtue of some such fact that minds, like thunderstorms or galaxies, would be real entities in good standing at all (regardless of the truth or otherwise of physicalism). But, even at best, the physicalist position is that any description of states and events, which incorporates *mental* states and events, will be exactly equivalent to some alternative (albeit vastly more complicated) description in terms purely of *physical* states and events. Indeed we might expect mentalistic descriptions to be mere approximations to the purely physicalist descriptions (though this is not crucial to the argument).[2]

In particular, consider any episode of the (apparent) exercise of "freedom"— that is, some kind of rational, or at least considered, decision making. If physicalism is true then, in principle, the initial set of mental states (and any other relevant factors) can be reduced to physical states; the trajectory of the system

[2] Both Smolensky (1991) (with his plea for the "Proper Treatment of Connectionism" or PTC), and Hofstadter (1979; 1983) (with his concept of "tangled hierarchies"), have given interesting discussions of such an *approximate* relationship between mentality and physics. A detailed review would take me too far afield here, but I briefly consider Hofstadter's views again in section 2.3.2 below.

can be evaluated by reference only to these physical states;[3] and the physical result or outcome (which, in the general, stochastically indeterministic, case will not be unique, but will rather be represented by a probability function or distribution) can then be encoded back into the resulting mental states, which will represent the decision (or a probability function or distribution over potential decisions).

This is, of course, simply a restatement of Laplace's thought experiment which envisaged a "demon" who could know the instantaneous dynamic state of the entire universe, and could therefore predict the entire behaviour of the universe for all future time. The only additional feature I have introduced is to allow for stochastic or probabilistic rather than strictly deterministic dynamics—in deference to the stochastic form of quantum mechanical physical theories. This does not, in any way, affect the force of the argument with regard to the exercise (or not) of human "freedom".

Note carefully that the argument does not rely at all on the *practicality* of a Laplacian demon. In particular, although Popper (1988) has provided a variety of arguments against what he terms " 'scientific' determinism", this latter doctrine is much stronger than the mere causal closure of the physical world claimed by H_p. " 'Scientific' determinism" seems to require that a Laplacian demon be physically realisable, at least in principle (I take this to be implicit in Popper's "principle of accountability" and his requirement for prediction from *within* the physical world). Whereas, in the discussion of "freedom" above the point is not whether the future can, in fact, be predicted (statistically or otherwise), but whether it can be *altered* (statistically or otherwise). In this respect, H_p is much closer to what Popper calls "metaphysical determinism", a doctrine implied by, but much weaker than, " 'scientific' determinism".

In any case, the essence of H_p is that no alternative analysis of the genesis of a human "decision", using mentalistic terms or otherwise, could say any *more* about the relationship of that "decision" to the prior state of the universe; indeed, we expect that a mentalistic analysis would yield, at best, only a poor, and

[3] There may be some difficulty with establishing what constitutes the "system" here; but, if needs must, we allow this to include the entire physical universe (regardless of whether this is bounded or unbounded). Recall that this is only an "in principle" discussion.

incomplete, approximation to the physicalist result. In short, mental states and events would have to be considered as, in some sense, *epiphenomenal.*[4] It is true that, under H_p, the outcome of a decision may not be deterministic (i.e. be a *unique* function of the prior state), but it cannot be reasonably said to be the "free choice" of the person; the possible outcomes, and their relative probabilities were already determined, and were not changed one iota by the particular thoughts that the person (appeared) to think.

The loss of freedom implied by H_p carries with it, of course, the loss of responsibility or moral obligation: since the person's thoughts (desires, intentions etc.) can be dispensed with in evaluating her actions, we could hardly hold her responsible for those actions.

Taken to its logical conclusion of course, this signifies that my very discussion of this topic is also epiphenomenal, and, in that sense, ridiculous (though perhaps not quite absurd). This result is, in essence, what Popper has termed "the nightmare of the physical determinist" (Popper 1965, p. 217) because it takes its clearest form under the hypothesis of a deterministic physical universe, in which a unique trajectory property holds. However, the point which Popper was at pains to expose is that *the nightmare is not in the least relieved by a stochastically indeterministic physics.* As long as a complete reduction of mental states and events to physical states and events is possible, in the sense that the resulting description is causally closed (whether the causation is deterministic or stochastic) then the nightmare recurs, as I have tried to make clear above. In Popper's words, "indeterminism is not enough" (Popper 1965; 1973).

2.3.2 Some Contrary Views

There exist, of course, a variety of contrary views on the repugnant consequences of physicalism.

Firstly, a common supposition is that stochastic indeterminacy (typically, though not necessarily, involving an appeal to quantum mechanics) can make physicalism and human freedom compatible. Indeed, this was the view of Arthur

[4] "Epiphenomenalism" comes in more than one flavour. The kind I have in mind here is that of Hofstadter (1983)—which seems to be subtly different from that of Popper (Popper & Eccles 1977, Chapter P3, Section 20).

Holly Compton, as noted by Popper in his Compton memorial lecture (Popper 1965); however, Popper firmly rejected this view, essentially for the reasons discussed in the previous section. The point is that physical indeterminacy of this sort simply does not change the nature of the argument, nor, therefore, its conclusion. I shall not consider this position further.

A.F. Huxley has argued that the *impracticability* of actually carrying out a physicalist reduction robs it of its sting:

> ... I used to be upset at the idea of possibly not having free will, but it now seems to me that even if we do not have free will, the events which govern our movements are so unpredictable that there is no need to be worried about it.
>
> Huxley (1983, p. 15)

Penrose has recently offered a more sophisticated variation on this argument, in the context of a discussion of free will. He argues that a complete physicalist analysis of any particular mental event or events may be not merely impractical, but actually *impossible*, in the technical sense of being *uncomputable* (Penrose 1990). In this way, the physical world could, in fact, be causally closed, but in such a way that this closedness could not be exploited from within.

In a sense, however, the relatively sophisticated appeal which Penrose makes to the notion of computability is unnecessary. The following argument, formulated by, for example, Popper (1974c, Section XXV), seems to me to establish the same point more directly and decisively. Reality, in its entirety, is causally interconnected (by definition). Thus, any *practical* attempt to make a complete analysis of any aspect of reality (from within the real world) would require a complete model of the real world to be embedded within itself; this would, of course, include a model of the model, and so on. This is clearly impossible (incompletable).

So let it be stipulated that a *complete* reduction of mentality to physics will always be impractical; the point remains that the repugnance of physicalism rests entirely on its *in principle* nature, and not any particular claim to be able to carry it out; the latter would be a factor in any attempt to *corroborate* physicalism, but that is not the issue just here. Indeed, Penrose himself seems to finally acknowledge that the impracticality of a physicalist reduction cannot, in itself, restore human freedom to the universe (Penrose 1990, pp. 558–559).

Another possible position is to accept the consequences of physicalism, but to put a brave face on the situation—claim that it may not be intrinsically repugnant after all.

Sperry put this view succinctly when he said "There may be worse fates than causal determinism" (Sperry 1965, p. 87). It should be stressed that Sperry does not mean *strict* determinism here—he specifically accepts that a stochastic *indeterminism* would add nothing more than a degree of "unpredictable caprice" to our actions. Rather, he is referring to the general physicalist position that mental events are ultimately reducible to physical events, which is to say H_p.

However, Sperry's position is still a good deal more complex than the slogan might suggest. Hofstadter (1985, Chapter 25) has provided an extended allegory expanding on this paper of Sperry's. Ultimately, in fact, both Sperry and Hofstadter seem to be ambivalent about the implications of physicalism for free will. That is, as far as I understand them, they adopt physicalism, accept that this is incompatible with "free will" as it is conventionally understood, and yet they also seem to qualify their physicalism, as if to draw back again from this abyss.

Thus, Sperry claims, in effect, that we can have our physicalist cake and eat it:

> ... you will note that the earlier basic distinction or dichotomy between mentalism and materialism is resolved in this interpretation, and the former polar differences with respect to human values ... become mainly errors of reductionism. This may be easily recognised as the old "nothing but" fallacy; that is, the tendency, in the present case, to reduce mind to nothing but brain mechanism, or thought to nothing but a flow of nerve impulses.
>
> ... Our quarrel is not with the objective approach but with the long accepted demand for exclusion of mental forces, psychic properties, and conscious qualities—what the physicist might class as "higher-order effects" or "co-operative effects"—from the objective scientific explanation.
>
> Sperry (1965)

Like Sperry, Hofstadter emphasises the existence of "emergent" behaviours, in the sense of levels of description having their own distinctive characteristics, even though these are still "compatible with" (but does this mean "reducible to"?) a purely physical level of description. In the case of conscious experience and free will, Hofstadter particularly emphasises the Gödelian implications of self referential symbols at different levels of description (Hofstadter 1979, Chapter XX).

It seems to me that both Sperry and Hofstadter are here confusing two quite different issues: the *utility* of mentalistic (or other "higher-order") descriptions, versus their *necessity.*

It is certainly the case that there exist descriptions of states and events in the world, incorporating mentalistic terms, which are approximately, if not exactly, true—indeed, it was stipulated in the previous section that this is actually the *defining condition* (at least in a causally closed physical world) for such "higher-order" entities to be recognised at all. These mentalistic descriptions, being "higher-order", are more concise and tractable than the corresponding purely physical descriptions. This is enough to make them useful additions to, or even replacements for, purely physical descriptions, for practical purposes of analysis and prediction.

But none of this implies that "higher-order" (mentalistic or otherwise) descriptions are *necessary.* Indeed, the point of saying that the physical world is causally closed is precisely to say that non-physical entities are, even if only in principle, superfluous to a complete account of physical states and events.

To be fair to Sperry and Hofstadter, neither explicitly claims that their approach does anything to restore the dignity of man in a soulless universe. At the end of the day, they are more concerned with reinterpreting our *attribution* of free will, than in restoring or rehabilitating the real thing. This is a perfectly sensible procedure upon the adoption of a physicalist position, but I do not see that it can make physicalism in the least degree more *palatable.*

There is a final possible position to be considered, though it really brings us full circle. This is to claim, *despite* the arguments marshalled in the previous section, that physicalism somehow *is* compatible with the exercise of free will, and the attribution of responsibility. This is a position which Dennett forthrightly promised to defend:

> ...can psychology support a vision of ourselves as moral agents, free to choose what we will do and responsible for our actions? Many have thought that materialism or mechanism or determinism ... threaten this vision, but ... *I consider the most persuasive of the arguments to this effect and reveal their flaws.*
>
> Dennett (1978b, Introduction, p. xxii, emphasis added)

However, virtually in the same breath, we already find a partial retreat from this bold and intriguing promise:

> By uncovering the missteps in the most compelling arguments for this thesis *I claim not to refute it, but at least to strip it of its influence.*
>
> Dennett (1978b, Introduction, p. xxii, emphasis added)

I shall ultimately find myself more or less in agreement with Dennett in this *second* formulation; that is, while I continue implacably to assert the repugnance of physicalism, I agree that this need not "influence" our scientific investigation of it. However, as I shall discuss in the next section, my grounds even for this circumscribed position are somewhat different from, and more general than, Dennett's.

But, before proceeding to that, I should like to comment briefly on the detailed arguments which Dennett actually presented (Dennett 1973).[5]

Dennett primarily argues for the validity of adopting what he calls the *intentional stance* toward certain systems, specifically including people. This is a necessary step in his argument since the intentional stance is, he says, "a precondition of *any* moral stance, and hence if it is jeopardized by any triumph of mechanism, the notion of moral responsibility is jeopardized in turn" (Dennett 1973, pp. 242–243).

This is all true, but is not, in my view, *germane.* There is no doubt that the intentional stance can usefully be adopted in many situations, and that this possibility is a requirement for intentional systems, like minds, to be recognised as such at all. But it is not clear that *anyone* is arguing to the contrary (i.e. to the effect that the ultimate *truth* of purely physical description would, in some sense, imply the *falsity* of mental, or intentional, descriptions). The point is not that mentalistic or, more generally, intentional, descriptions are false, or even useless, but rather that they may be causally redundant. The physical world might, as it were, go along just the same way without them.

The notion that physicalism might somehow rule out the adoption of the intentional stance, for *utilitarian* purposes, is a distraction—a mere straw man. That

[5] Dennett has since provided a much more extensive analysis of "free will" and related problems (Dennett 1984). Chapter 5 of that work addresses the issues of most concern for my purposes, but I have been unable to identify anything which would deflect the criticisms which I present of Dennett's earlier essay (Dennett 1973). A properly comprehensive review of (Dennett 1984) would take me too far afield; I shall therefore not discuss it further.

is, as long as the intentional stance is merely that—a “stance” we might choose to take up with respect to certain physical systems, for utilitarian purposes—it seems that it cannot be relevant to the issue under discussion here.

However, Dennett does offer a few further twists that might affect this conclusion. He considers the point that to abandon the intentional stance toward *oneself* would be fundamentally incoherent; that there is therefore an element of intentionality in the world which *is* more than an optional stance toward an essentially physical system—for the very taking up of a stance is, in itself, an intentional action.

This is an intricate and intriguing argument. But Dennett himself immediately admits that it is really an attempt to refute physicalism, rather than a means of reconciling physicalism with free will. And, as a refutation of physicalism, it fails. Briefly, it is another deterministic nightmare: if physicalism is true, we cannot properly be said to “choose” to take up any stance at all.

Popper has discussed this kind of argument critically, and provides the following concise version of what can, and cannot, be validly drawn from it:

> ... the epiphenomenalist argument leads to the recognition of its own irrelevance. This does not refute epiphenomenalism. It merely means that if epiphenomenalism is true, we cannot take seriously as a reason or argument whatever is said in its support.
>
> Popper & Eccles (1977, Chapter P3, p. 75)

Indeed, I may say that this analysis provides the rationale for the entire orientation of the current chapter: I consider the arguments *against* physicalism, but not those *in favour*; for, by definition, the most compelling arguments in favour of physicalism must also be the most self defeating.

But to return to Dennett, he next considers the point that:

> ... no information system can carry a complete true representation of itself ... And so I cannot even in principle have all the data from which to predict (from any stance) my own future.
>
> Dennett (1973, p. 254)

But this is simply back to the question of the *practicality* rather than the *truth* of physicalism; indeed, Dennett explicitly acknowledges Popper’s formulation of this point, as I have already described it in the discussion of Huxley and Penrose above; and it still does not impinge on the issue of free will.

It seems then that Dennett does not achieve his original aim of showing how free will and physicalism might genuinely co-exist in a single cosmology. His concluding remarks are, in fact, addressed to a different theme:

> Wholesale abandonment of the intentional is in any case a less pressing concern than partial erosion of the intentional domain, an eventuality against which there can be no conceptual guarantees at all.
>
> Dennett (1973, p. 255)

The issue is no longer the relationship between free will and physicalism; but rather the potential for abuse of whatever physicalist understanding of mentality (if any) may, in practice, be achieved. This is now a discussion of the *uses* of science, which is to say a *moral* discussion. As such, it is not, itself, any longer a part of the scientific discourse. This is the point at which I can finally agree with Dennett, and I elaborate this general position in the next section.

2.3.3 But: does it *really* matter?

> ...Thus I regard the doctrine that men are machines not only as mistaken, but as prone to undermine a humanist ethics. However, this very reason makes it all the more necessary to stress that the great defenders of that doctrine—the great materialist philosophers—were, nevertheless, almost all upholders of humanist ethics. From Democritus and Lucretius to Herbert Feigl and Anthony Quinton, materialist philosophers have usually been humanists and fighters for freedom and enlightenment; and, sad to say, their opponents have sometimes been the opposite.
>
> Popper & Eccles (1977, Chapter P1, p. 5)

Does it matter that the physicalist hypothesis has dehumanising implications? Well, the fear expressed above by Popper, that it might be used as an excuse for dehumanising *actions*, is not entirely without foundation. Perhaps this explains, in part, why a proponent of physicalism might be loath to accept that this position does indeed imply the abandonment of human freedom and responsibility. If so, this would be quite understandable, perhaps even admirable in its way; but I suggest that it would also be quite mistaken.

We can and should face up to the consequences of our theories, *even* when they are odious. We can do this because, in fact, there is nothing to fear from even the most odious consequences of any theory—*provided we remember that our theories are just that, fallible inventions of the human mind.* I cannot accept that any such fallible theory, no matter how well corroborated, could ever provide us with *moral* principles or, worse, *justifications*. "Scientific morality" is, I suggest, a contradiction in terms. As Popper has said, even the greatest defenders of physicalism have actually been upholders of the humanist ethic—and that, in my view, is precisely as it should be.

In short, I assert that it is only good science to admit the implications of our theories, repugnant or not; *but that it is then only good philosophy to admit that our scientific theories, in themselves, are devoid of moral authority.* Science absolves no sins.

In this present context, this means that the implications of physicalism, repugnant as they may be, can still be viewed with a certain degree of equanimity or detachment. We might almost say that this makes the attribution of repugnance aesthetic rather than scientific; as such, it need not, and should not, deflect the scientific investigation.

Having said that, I should emphasise that I do not suggest that the path of science is free from moral decisions, or from moral culpability—that scientific "progress" might be justified as an end in itself. Quite to the contrary, I consider that scientific activities are no different from any other human activities in this respect; they share the moral imperative for us to consider (as well as possible), and accept responsibility for, the likely outcomes of our activities. It is precisely in discharge of this moral obligation that I have stipulated my abhorrence of physicalism, *per se*, have positively argued that this abhorrence is justified, but have then gone on to argue that this, in itself, does not have the force of a general, moral, restraint on the scientific investigation of physicalist theories of mentality.

This position should be distinguished from, say, a specific advocation of scientific investigation into theories of "brain-washing, subliminal advertising, hypnotism and even psychotherapy ... and the more direct physical tampering with drugs and surgical intervention" (Dennett 1973, p. 255). Such activities could, no doubt, fall within a physicalist research programme; but, as Dennett implicitly draws out, they would require specific moral validation well beyond anything which has been discussed here.

2.4 Refuting Computationalism?

Having made clear the unhappy implications of H_p (and thus H_c), but having also affirmed that this should not, in itself, deflect us from the further scientific study of these hypotheses, I now return to the substantive question of this chapter: *Has H_c already, in fact, been refuted?*

One avenue for the attempted refutation of H_c is the claim, originally propounded by Lucas (1961), that Gödel's results on the existence of undecidable propositions in consistent formal systems establish that mentality is necessarily irreducible to formal processes. However, this has already received extensive exploration, and many detailed criticisms (see, for example, Hofstadter 1979; Dennett 1970; Hofstadter & Dennett 1981, p. 470, p. 475 give further references). I shall therefore make only one brief comment here.

The Lucas argument relies on the claim that, faced with any machine which putatively exhibits mentality, one can always formulate a proposition which the machine cannot prove but which any *person* (Lucas himself, for example) can see to be "true". As Dodd (1991) has pointed out, albeit in a slightly different context, any such perception of "truth" is actually dependent on an assumption of *consistency* for the relevant formal system; but, precisely because of Gödel's results, such consistency *cannot*, in general, be proven. Thus, it seems to me that the argument by Lucas fails from the very start: while the machine cannot prove its Gödel sentence, *neither can Lucas*; the most that Lucas can do is to *conjecture* that it is true—and I can see no bar to the machine also doing that much. In any case, I shall not pursue the Lucas argument further.

I now turn to two other, quite distinct, arguments for the refutation 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.

2.4.1 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 twelve 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). It is clearly a matter of some continuing interest and significance for AI, and I should therefore like to comment on it.

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 2.4.2 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 another variant of the "deterministic nightmare" of section 2.3.1). 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 entirely 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, in essence, to Drew McDermott, who introduced it in personal communication with Harnad; I have not identified any published version of precisely this idea. In my view, this argument is not only concise and elegant, but also 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).[6]

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.

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 can 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.

[6] 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`.

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.

2.4.2 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, operationally, *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). 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.

2.4.2.1 Criticism by Dennett

Dennett has provided a more or less detailed criticism of the position of Popper & Eccles, in the form of a book review (Dennett 1979). In Dennett's own words, this is a "caustic" review (Hofstadter & Dennett 1981, p. 477), where he finds very little of any sort to approve of, and appears to consider the arguments to be at best flawed, and at worst incoherent.

If Dennett were successful in his criticism, there would be nothing further for me to say here. However, while I generally agree with his conclusions, I consider that his route to them is quite inadequate, so there is still some work for me to do.

This inadequacy is presumably partly due to the constraints of the book review format. However, this cannot excuse, for example, Dennett's parenthetical summarising of Popper's *World 3* as "essentially a platonic world of abstract entities, such as theories, hypotheses, undiscovered mathematical theorems" (Dennett 1979, p. 94). The superficiality of this comment should be clear when it is noted that Popper actually expends several pages of argument to *distinguish* his World 3 from Plato's world of ideals (Popper & Eccles 1977, Chapter P2, Section 13).

Or again, Dennett severely criticises the apparent incompleteness of Popper's position:

> What kind of interaction can this be between a thinking and a theory? We are not told. Popper waves his hands at how modern physics has vacated all the old-fashioned philosophical ideas about causation, but does not give a positive account of this new kind of causation...
>
> Dennett (1979, p. 94)

But, when Eccles attempts to provide some analysis precisely of the nature of this causation, Dennett indulges his sarcasm from the opposite direction, accusing Eccles, in turn, of incompleteness because he:

> ... passes the buck to "the self-conscious mind," about whose apparently wonderful powers he is conveniently silent.
>
> Dennett (1979, p. 95)

Thus Dennett has managed to criticise each author for not covering issues dealt with by the other, and all this after peremptorily stating, in his introductory remarks, that:

> These men are not really co-authors, but co-contributors to an unedited anthology; they have not hammered out a joint theory, nor does it appear that they have been tough critics of each other's contributions.
>
> Dennett (1979, p. 92)

It seems that Dennett's review might have benefited from some tough criticism itself.

In summary then, while I agree with Dennett that the arguments propounded by Popper & Eccles are flawed, I consider that he has failed to confront them with the seriousness which they demand; and that, even where his criticism is well-founded, its credibility is undermined by embellishments which are not necessary, nor even consistent.

2.4.2.2 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.[7]

2.4.2.3 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, p. 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".

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.

[7] 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.

2.4.2.4 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 rebuttal will therefore be similarly twofold.

2.4.2.5 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 *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* a necessary 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.

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.

2.4.2.6 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.

This defect in Popper's argument is, indeed, pointed out by Dennett in his review Dennett (1979). However, his presentation is somewhat simplistic, if not actually mistaken—as when he says:

> It seems just as apt to say that when I put a Z brace on a gate to keep it from sagging, I bring about a causal interaction between theorems of Euclid and the pine boards, as it does to say that there is a causal interaction between my thinking and these theorems. That is, in the absence of much more detailed persuasions, both views appear ludicrous.
>
> Dennett (1979, p. 94)

It seems that Dennett is presenting here, not an argument as such, but an example of the kind of thing which he himself has described (in a different context) as an *intuition pump*—i.e. a thing which is "not, typically, an engine of discovery, but a persuader or pedagogical tool—a way of getting people to see things *your* way once you've seen the truth" (Dennett 1980).

I suggest that, in fact, the attempted *reductio ad absurdum* fails to fully confront Popper's argument. For the *crux* of Popper's argument is not that World 3 entities, *in general*, can only interact with World 1 via World 2 (though he does, admittedly, claim this); the important point is the much more particular claim that this is so for certain *specific* World 3 entities, namely those which are "unembodied" or demonstrably irreducible to World 1 entities. For Dennett to properly refute Popper's argument, his example of a direct interaction between World 3 and World 1 would have had to involve some such *unembodied* World 3 entity, rather than just any arbitrary World 3 entity. His example is not of this sort; or at least, is not *clearly* so. I shall return to this below.

Thus, while I have restated Dennett's point—that Popper has failed to establish that unembodied World 3 entities *cannot* interact directly with World 1—I have not relied for this on Dennett's suggestion that unembodied World 3 entities can, in fact, interact directly with simple unproblematic World 1 entities such as pine boards. Rather, I am willing to stipulate, with Popper, that the interaction requires the mediation of World 2; but then point out that this observation, in itself, is neutral with respect to the reducibility of World 2, *in some fashion,* to World 1.

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 (this would include, for example, Dennett's pine gate), 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. In short, his argument fails because it is ultimately circular.

However, whether one accepts Dennett's simplified analysis, or insists upon the more detailed refutation presented here, the 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. From both Dennett's point of view and my own, this is, arguably *enough*: we have provided a sufficient basis to refute Popper's argument for mind-body dualism, which is all we really sought to do; and, indeed, that is where Dennett does leave the issue. 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.

This example may be said to refine and elaborate the earlier example suggested by Dennett, of an interaction between theorems of Euclid and a Z brace. It goes beyond Dennett's example just insofar as it stipulates that, in the initial state of the physical system, the theorem in question is only implicit in a set of axioms—and is thus "unembodied" in Popper's sense—and yet "interacts" with the system in just the sense with which it might be said to interact with a mathematician.

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; 1970b), 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.

2.5 Conclusion

In summary, my claims in this chapter are:

1. That physicalism in general, and computationalism in particular, are irredeemably repugnant to human values and to the dignity of mankind; it seems to me craven to deny this.

2. That neither physicalism nor computationalism have (yet) been definitively refuted; in particular, two distinct kinds of argument, by Searle and Popper respectively, purporting to achieve such a refutation, are flawed.

The relationship between these two points is, I think, very important. It seems to me that the first point 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, in Chapter 1, I 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 (1974c, 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. For the remainder of the Thesis then, I shall tentatively *adopt* this computationalist thesis, H_c, and explore at least some of its detailed ramifications.