THE INTROSPECTION GAME

Or, Does The Tin Man Have A Heart?

Andrew Clifton

Abstract

Eliminative functionalism is the view that mental attributes, of humans and other machines, consist ultimately in behavioural abilities or dispositions. Hence, 'Strong AI': if a machine consistently *acts* as if it were fully conscious, then conscious it is. From these assumptions, optimistic futurists have derived a variety of remarkable visions of our 'post-human' future; from widely-recognised 'robot rights' to 'mind uploading', immortality, 'apotheosis' and beyond. It is argued here, however, that eliminative functionalism is false; for at least on our present knowledge, the subjectively qualitative characteristics of conscious experience are neither deducible from, nor logically *required* to generate, the performance of any sort of overtly 'intelligent', or indeed, characteristically human behaviour. Thus, a machine could easily be designed to *report* awareness of phenomenal qualities, without necessarily possessing them; and Alan Turing's 'Imitation Game' test for artificial *thinking* is unable to determine whether or not a machine is *sentient*. An alternative test is proposed, in which the machine is asked phenomenological questions under conditions designed to detect any form of cheating—whilst also, potentially revealing evidence for the occurrence of genuine qualitative experience.

Keywords

Artificial intelligence, consciousness, extropianism, functionalism, phenomenal qualities, post-humanism, qualia, sentience, Strong AI, Turing test

1 INTRODUCTION

The 'Strong AI'[1] conception of the mind is commonly taken to assert not only that a suitably programmed computer will be intelligent, in precisely the same sense that a human being is intelligent, but also that the same claim may be made for all other mental attributes. Thus it is possible, at least in principle, for such a machine to experience emotion, qualitative sensation, pleasure, pain, aspiration, compassion, sorrow—indeed, all of those special aspects of conscious experience that we, as humans, consider valuable, important and essential to our nature. A further central tenet of Strong AI consists in the assumption that overt behaviour provides satisfactory evidence for the attribution of mental states. As the artificial intelligence pioneer John McCarthy stipulates: "A sufficient reason to ascribe a mental quality is that it accounts for behaviour to a sufficient degree." (McCarthy, 1995). On this view, therefore, the 'Imitation Game', proposed by Alan Turing (1950) as an objective criterion for resolving the question, "*Can machines think*?", may with equal validity be applied to the question "*Can machines feel?*"; and in the case of any computer or robot which consistently passes such a test, we may confidently affirm, so to speak, that the Tin Man has a heart.

The Imitation Game, as Turing introduces it, initially involves no computer at all, but rather three human beings: a man (A), a woman (B) and a (male or female) interrogator (C)—whose task it is to correctly identified the gender of (A) and (B). These three participants are situated in separate rooms and communicate, anonymously, via typewritten messages. (B) is required to be honest and helpful throughout, whereas (A)'s objective is to convince the interrogator that he, (A), is female. The question, at this stage, is simply: can the interrogator determine which

[1] A term originally coined by John Searle; see Searle (1980)

of the two putative 'women' is, in fact, the female impersonator? Next, we imagine a variation on this theme—as Turing explains:

> We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?" (Turing, 1950, p 443).

Turing suggests that if a machine's responses in the imitation game are indistinguishable from those of a human being, then we may say that the machine thinks. He is not, as some critics have assumed, setting out to *define* intelligence in such operational terms, but rather to propose an operational test as a *sufficient* inductive indication of its presence.[2] He goes on to argue that a machine capable of passing such a test is possible, in principle.

Recent debates upon the merits of these claims have focused, primarily, on intellectual faculties; asking, for example, (i) whether various mental capabilities, such as creativity and mathematical insight, can be implemented algorithmically (Lucas, 1961, 1996; Penrose, 1989, 1994); (ii) whether a purely algorithmic system can plausibly be said to possess intentional states like 'knowing' and 'understanding' (Searle 1980, 1984, 1992, 1997) and (iii) whether or not intelligence can be simulated (Block, 1982).[3] In what follows, however, I shall assume, for the sake of argument, that Turing is right in two important respects: first, that *at least* in so far as 'thinking' is characterised by a capacity to *behave* intelligently, the Imitation Game is a plausibly sufficient empirical criterion for its occurrence; second, that it is possible, at least in principle, to construct a discrete state machine that is capable of passing this test.

There remain, however, two further important questions which seem, hitherto, to have been relatively neglected. First: could a computer or robot, endowed, let us assume, with a more-or-less 'human' level of operational intelligence, also possess subjectively *qualitative* experiences, which are, at least in some general sense, comparable to our own? In other words, could it experience a phenomenal world of sensory experiences—subjective colours, tastes, sounds, smells and so on—which is, at least *roughly*, similar to ours? Second: could such a machine also possess those properties of conscious experience which give their owner a *valuative* sensibility, such that the quantity and quality of its conscious experience *matters*—at least, to itself and arguably, on ethical grounds, to ourselves? Could it, in other words, be said to have *interests*, in roughly the same way that we do—and thereby qualify, in a morally relevant sense, as a *person*?

In respect of these two questions, we may now add a third: if a machine displays, with unfailing accuracy, every kind of overt behaviour which is typically associated, in humans, with the above mental attributes, are we justified in assuming that it necessarily possesses them? Consider the imagined predicament of the Tin Woodman in Frank L Baum's *The Wonderful Wizard of Oz*—who complains that, having no heart, he lacks the capacity for love. We observe that, nevertheless, he *behaves* just as if he experiences emotion and feels compassion for others.[4] Should we conclude that such *behavioural dispositions* constitute all that really matters about the 'heart' whose absence the Woodman, mistakenly, regrets?

For supporters of Strong AI, the answers to all of the above questions must be resolutely affirmative. The standard defence of this position is quite straightforward, if we assume—as do many contemporary philosophers and cognitive scientists—that all mental attributes may be

---

[2] On this point, see Moor (1976, 1987, 2000, 2001).

[3] For critical discussions of these three major arguments, see e.g. Boden (1990) and (respectively) (i) Hadley (1987), Grush & Churchland (1995), Chalmers, (1996b) (ii) Preston & Bishop (2002) (ii) Dennett (1985, 1987, ch. 9)

[4] Baum, (1900). Likewise from time to time, the Scarecrow behaves as if he has intelligence and the Lion as if he has courage. Baum's point, of course, is directed at the human reader: we should judge other—and indeed ourselves—on the basis of outward behaviour rather than physical characteristics.

fully defined in terms of functional patterns of objectivity observable causes and effects. On this view, we must *eliminate* from our ontology all 'folk psychology' notions of mental properties which cannot be described in such terms. It follows that mental states have the property of *multiple realisability*; their functions are independent of the physical mechanisms which sustain them. Therefore, the presence of any mental attribute is sufficiently evidenced by appropriate patterns of observable behaviour; in other words, the Turing Test is valid *as a test for sentience* (as opposed to mere operational intelligence) and Strong AI is true.

With this rationale, Strong AI has established itself as a popular assumption not only within the various academic sub-disciplines of cognitive science and the philosophy of mind, but also amongst popular science-fiction writers and optimistic futurists—particularly those who count themselves 'post-humanists' or 'extropians'. Combining Strong AI with a number of more-or-less plausible assumptions regarding the achievable potential and pace of progress within the relevant technological fields, various authors have argued that:

(1) in the fairly near future (i.e., within a few decades) we will be able to form close and meaningful emotional relationships with *fully sentient* artificial companions (Kurtzweill 1995; Sloman, 2000; Grand, 2001);
(2) for any computer or robot to display intelligence comparable to our own, it will almost certainly *require* qualitative and valuative experience, including emotions (Minksy 1986; Sloman and Croucher, 1981; McDermott, 2001); indeed, sufficiently intelligent machines might even be expected to have spiritual experiences and develop religious beliefs (Furze, 1995);
(3) on ethical grounds, we ought to recognise and respect the inalienable rights of intelligent machines—and actively resist any discriminatory efforts to restrict their autonomy and limit their capabilities (Elton, 1997, 2001; Holst, 2001; Mcnally and Inayatullah 1988; Inayatullah 1998; Tipler, 1995);
(4) it will eventually become possible to 'upload' or transfer human consciousness into a computer or robot, thereby effectively attaining immortality (Moravec 1990, 2000; Paul and Cox, 1996; Broderick, 1999, 2001);
(5) superstitious resistance to this technological metamorphosis will be both foolish and futile; since economic competition and/or violent conflict between super-intelligent machines and any remaining flesh-and-blood humans will render the latter obsolete—and most probably, extinct (Moravec 1990, Vinge, 1993);
(6) as immortal post-humans, we will be able to colonise the galaxy (and ultimately, the entire universe) by means of a geometrically expanding migration of self-replicating spacecraft—'Von-Neumann Probes'—which systematically convert all of the usable materials they encounter into duplicates of themselves, to be despatched elsewhere (Tipler, 1995);
(7) all of the above is a consummation devoutly to be wished—for, in post-human, silicon-form, we will ultimately achieve *apotheosis*—our intelligence and capacity for conscious experience expanding exponentially, until we become, by any reasonable criterion, *gods* (Yudkowsky, 2001).

Of course, such grandiose conceptions of possible futures for humanity all depend heavily upon the assumption that the foregoing argument for Strong AI is sound. In this essay I will argue, on the contrary, that at least with respect to the questions which concern us here, the strict, eliminative functionalism upon which Strong AI depends is false. It is not my intention to claim that the functionalist approach is wholly without merit; but rather, that it is *insufficient*. Functional definitions may usefully serve to *individuate* mental attributes, but they do not always adequately describe them. It follows that both the principle of multiple realisability and hence, Strong AI are also false; thus, the Turing Test is utterly unable, in principle, to demonstrate that a particular machine is sentient—in the sense in which we attribute sentience to ourselves.

This is not to say that *no* artificial system *could possibly* possess qualitative and valuative mental attributes, but rather that the possession of them is neither a necessary consequence of any sort of intelligent functioning, nor infallibly revealed by displays of the sort of overt behaviour with which they are commonly associated in human beings. This is because, regardless of the metaphysical position we favour with respect to the fundamental nature of consciousness, it seems evident that—whatever their nature—the qualitative and valuative features of experience are *functionally redundant* with respect to cognitive and behavioural performance. I am not suggesting here that they have *no* functional role, but rather, that their peculiar characteristics are not essential to the roles they appear to serve; they are, so to speak, surplus to requirements. Thus, until we are able to determine exactly what these states are—and how they come to arise in our experience, as a consequence of brain activity—it is, *prima facie*, highly unlikely that any artificial systems we create, *based upon any cybernetic principles that we current understand*, will, fortuitously, just happen to possess them.

If these arguments are valid and their premises true, it follows that the extropian visions of the future outlined above represent not so much a realistically attainable brave new world as a fool's paradise. Rather than leading us onward to immortality, widespread faith in their validity may well have extremely undesirable consequences—quite possibly, on a hitherto unprecedented and utterly catastrophic scale.[5]

The remainder of this paper is structured as follows. In §2 I will examine the origins, motivation and supposed justification of the version of functionalism which supports Strong AI. In §3 I introduce a definition of 'phenomenal qualities' and argue that the existence of these features of conscious experience is incompatible with Turing test functionalism. In §4 I present my functional redundancy argument against the validity of the Turing test as a criterion for sentience. In §5 I defend the 'Moderate AI' view that sentient machines are possible in principle—but argue that until we are able to solve the mind-body problem, they are unlikely to be created in practice. In §6 I examine the possibility of replacing the Turing Test with an alternative investigative approach to assessing the putative sentience of artificial information systems. I will call my proposed technique 'The Introspection Game'.

2 FUNCTIONALISM AND STRONG AI

The earliest origins of functionalism may be traced, perhaps, to such varied formative influences as the scepticism of David Hume, the positivism of August Comte and the evolutionism of Charles Darwin and Herbert Spencer; yet it seems to have found its first definitive expression in the pragmatism of Charles S. Peirce. The essence of this paradigmatically modern position is clearly expressed in the famous Peircian maxim: "Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object." (Peirce, 1878). If we apply this epistemic precept to our understanding of the mind, then we will say that to have a mental faculty, or to experience a mental state, is nothing more than to be *disposed to do something*; hence, performance of that function implies the existence of the corresponding mental attribute. On this view, therefore, mental events should be interpreted, at least primarily, as mechanisms which serve the organism in adapting to its environment; what consciousness contains is less important than what it does. Founded upon this principle, the functionalist (or 'instrumentalist'[6]) tradition within psychology and the social sciences was developed and popularised by a succession of American pragmatists, notably William James, John Dewey and George Herbert Mead (see e.g. Murphy, 1990; Shook, *ed.*, 2000).

Pragmatic functionalism provided a fertile intellectual background for the development of a family of related ideas in psychology and philosophy, all of which share a common

[5] For further discussion of some of these hazards and how we might avoid them, see Clifton (2004) [c] .

[6] Dewey's preferred term for his functionalist philosophy.

perspective which may be described, I suggest, as '*third-person absolutism*'. This widely (if often, tacitly) supported ideology combines strict anti-essentialism (i.e., opposition to the idea of *essences* or 'things-in-themselves') with the ultimate aim of achieving perfectly objective knowledge by banishing all trace of subjectivity and introspection from the respectable conduct of scientific inquiry.[7]

Methodological Behaviourism, for example, strictly forbids explanatory reference to mental properties; treats the organism as a "black box" and seeks to explain behaviour solely in terms of relations between stimulus and response (Watson, 1912; Skinner, 1953). *Operationalism* holds that theoretical concepts, such as mental states, must be defined solely by the techniques used to measure them (Bridgman, 1927). *Verificationism* says, similarly, that the meaning of a scientific statement consists only in its methods of confirmation (Ayer, 1935; Reichenbach, 1953; Schlick, 1959). *Emotivism* declares that ethical statements (e.g., "the Nazis were *really* bad!"[8]) are expressions of behavioural or emotional dispositions; they have no independent truth value and do not express knowledge (Ayer, 1935; Stevenson, 1937, 1944). Although widely criticised of late (see e.g. Brink, 1989; Sayre McCord, *ed*., 1988; Pojman, 1994; Audi, 1997), emotivism can still claim a number of eminent academic supporters (e.g. Harman 1977; Mackie 1977; Williams, 1973, 1982, 1986). Its various philosophical fellow-travellers, however, are now commonly acknowledged to be seriously flawed, in a number of different ways; hence, they are seldom explicitly defended today.[9]

In the middle of the last century, however, when these doctrines were still in vogue, they contributed greatly—at least, in spirit—to the development of two closely-related traditions which still enjoy a lively popularity: *logical behaviourism* and *eliminative materialism.* The former holds that mental states may be defined as dispositions to behave in particular ways, under particular conditions; the latter, that alleged mental items which fail to be captured by such a definition do not, in fact, exist. Important versions of these twin-doctrines were notably expounded (in their own distinctive styles) by Wittgenstein (1953) and Sellars (1956). Their clearest and perhaps most influential expression, however, may be found in the works of Ryle (1949) and Quine (1953); whose student, Daniel Dennett,[10] is prominent among those who continue to defend them—in the modified form which we recognise today as *eliminative functionalism*.

This is, of course, the thoroughly modern kind of functionalism which is widely considered to support Strong AI. While careful to avoid at least some of the weaknesses of its philosophical forebears, it maintains much of the tradition of positivism and behaviourism, as outlined above—remaining committed to the underlying perspective of third-person absolutism. Indeed, this basic outlook constitutes its supposed justification—for on this view, functionalism is a necessary consequence of a strictly scientific worldview.

There is no denying the popularity of this position. With only minor variations, eliminative functionalism is expounded, defended, or merely taken for granted by such distinguished philosophers as David Armstrong, Paul and Patricia Churchland, Daniel Dennett, Jerry Fodor, Gilbert Harman, Douglas Hofstadter, David Lewis, William Lycan, Hilary Putnam, Georges Rey, Richard Rorty and many more besides. It has also been defended by such highly accomplished cognitive psychologists as Richard Gregory, Philip Johnson-Laird and David Marr; and of

---

7 The futility of this enterprise is meticulously argued by Polanyi (1958, 1966). See also Clifton (2004) [a]).

8 Philip Gasper, quoted in Sayre-McCord (1988), p.1

9 For landmark criticisms of these ideas, see Chomsky (1959), Putnam (1962, 1975a, 1975b), Quine (1951). For current revivalism, see Misak (1995), Friedman (1999) and Zuriff (1985).

10 As Dennett has commented, in interview: "It's clear to me now that both Ryle and Quine had an enormous influence on my work. Maybe I'm what you get when you cross Ryle with Quine and add a little neuroscience." Pyle (1999).

course, by prominent artificial intelligence researchers, such as Rodney Brooks, John McCarthy, Drew McDermott, Marvin Minsky, Aaron Sloman, John Taylor and numerous others.[11]

It is not the case, however, that every conceivable type of functionalism consistently supports both Strong AI and the validity of the Turing test as a criterion for sentience. In order to do so, I suggest, 'Turing test functionalism'[12] must affirm the following thesis:

> *All mental attributes may be fully defined by some set of causes and effects which: (a) are objectively observable or measurable, directly or indirectly—and (b) satisfy some independent, objectively specifiable criterion or purpose.*

A number of further observations may help to clarify the implications of this claim.

(1). This strong version of functionalism incorporates the eliminative thesis that there are no mental properties which cannot be captured in a functional definition; hence, we may be properly call this position *eliminative functionalism*. It does not, as some critics suggest, deny that there are *any* first-person facts. Rather, it claims that all *legitimate* first-person facts must be, *somehow*, implicit in the totality of third-person facts—thus, once all of the latter are known, nothing further will remain to be explained.

(2). The sort of functional criterion which suffices to *define* a mental attribute or faculty is normally taken to be a *proximal* or particular one, e.g., 'learning to talk' or 'recognising faces'; as opposed to broadly *explanatory* functions, such as evolutionary value, which might be thought to account for the origin and existence of a great many *different* functional phenomena.

(3). The set of causes and effects which pick out a given mental state may legitimately include other (functionally defined) mental states; thus, functionalism has the merit of acknowledging all the information-processing complexities of conscious cognition, in accordance with the findings of cognitive psychology.

(4). However, the requirement that such defining causes and effects must be, in principle, *objectively observable* preserves a strong behaviourist slant. Moreover, if we accept the plausible assumption that there is no objectively practical or evolutionary utility in endlessly covert interactions between internal states, then the functionality of all mental states must ultimately emerge in speech or action. Thus for contemporary functionalists, as Paul Churchland avows: "The essential or defining feature of any type of mental states is the set of causal relations it bears to… bodily behaviour." (Churchland, 1984, p. 36) We may, therefore, describe this view as a form of *behavioural* functionalism.

(5). Eliminative, behavioural functionalism implies that mental states may be instantiated in many different ways. This is a consequence of the application, to mental states, of the principle of *multiple realisability*—which holds that whenever it is physically possible for a given functional role to be implemented in different mechanical ways, these alternative realisations are equivalent.[13] To cite an oft-quoted example, the functional role of a mousetrap may be fulfilled by a wide variety of different devices,

---

[11] See, e.g.: Armstrong (1968), Churchland, P.M. (1984, 1996), Churchland P.S. (1986, 2002), Churchland P. M. and Churchland P. S. (1998), Dennett (1978, 1987, 1991), Fodor (1975, 1980, 1981, 1990), Harman (1999), Hofstadter (1979), Hofstadter & Dennett, (1981), Lewis (1980), Lycan (1987, 1996), Rey (1980, 1996, 1997a, 1997b), Rorty (1972), Coterill (1997, 1989), Crick (1994), Crick & Koch (1990); Gregory (1987), Johnson-Laird (1987, 1988), Marr (1981), Brooks (2003), McCarthy (1995), McDermott (2001), Minsky (1986), Sloman (1978, 2000), Taylor (1995, 1999).

[12] Not to be confused with Putnam's (1975) 'Turing *machine* functionalism'. Other varieties of functionalism —e.g., the causal functionalism of Armstrong (1968) and Lewis (1980); the hierarchical or 'pandaemonium' functionalism advocated by Dennett (1991)—may all qualify as versions of turing test functionalism, on the definition specified here.

[13] This principle also argues against the coherence of a functional definition of a mental state in terms of details of any particular physical instantiation of the system—thereby reinforcing the behaviourist stance of Turing test functionalism.

made of different materials and operating on different principles—yet they all *count*, with equal validity, as mousetraps. Similarly, it is argued, any given mental state can be realised by a mechanism which successfully reproduces the same defining pattern of behaviour, in response to relevant stimulation, which individuates it in the case of human beings. In this way, of course, functionalism is considered to imply Strong AI—and the objective validity of the Turing test as a criterion for sentience.

Notwithstanding its widespread acceptance, some critics have challenged the view that eliminative functionalism provides unambiguous support for Strong AI. Strong AI may be understood to claim that it is impossible to 'simulate' mental attributes without actually *having* them; hence, *if you* (successfully) *fake it, you make it.* While it may seem plausible that this principle applies to abilities, it is surely open to question in the case of dispositions, where behavioural criteria are more ambiguous. These considerations give rise to what might be called the '*lying game*' argument—against Strong AI and other doctrines within the functionalist tradition (see e.g., Chisholm, 1957; Geach 1957; Putnam, 1975; Block 1979). If we notice that our casual acquaintance, Jack, consistently solves crossword puzzles without assistance, we may safely agree that he possesses the relevant cognitive ability; but if, just as frequently, we hear him saying, "I love Jill", we may not be quite so confident in attributing amorous feelings. A given pattern of behaviour, such as uttering declarations of affection, may be the product of more than one sort of mental disposition. When it comes to feelings, beliefs and emotional attitudes, people often *lie*, or affect a show of emotion which they do not feel.

A striking example of this phenomenon is provided by those people, colloquially known as 'psychopaths', who are clinically diagnosed today as having 'Antisocial Personality Disorder' or 'APD' (American Psychiatric Association, 1994). It has been widely observed that such people seem to be virtually incapable of feeling sympathy and compassion; indeed, it appears that in some cases at least, they literally *do not know* 'what it is like' to feel genuine concern for another's well-being (Reid et. al, 1986). However, psychopaths are often skilled and convincing actors; they are adept at *faking* normal human sympathies and hence, they often excel as professional con-artists (and also, no doubt, as lawyers and politicians). Imagine now the predicament of a parole board, interviewing an apparent 'model prisoner' who is, in fact, a un-diagnosed case of well-disguised, but extreme, APD. In this situation, the board-members are in a similar position to that of the interrogator in a Turing test. They don't know, *a priori*, whether the prisoner is a (more-or-less) normal person, guilty of some tragic crime of passion, but genuinely remorseful and reformed in character—or an irredeemable psychopath, covertly planning a spectacular new killing-spree. Most importantly, given sufficient acting ability, the prisoner's *observable behaviour* during the interview will provide no reliable evidence either way. Likewise, it seems plausible that a computer might be programmed to *cheat* in the Turing test; demonstrating nothing more than a plausible *facsimile* of emotion, feeling and sentience.

For eliminative functionalists, however, this objection is spurious. The point of the Turing test is not to determine *exactly what* a computer or robot thinks, believes or feels, but rather to establish whether or not it possesses *any* such mental attributes. Functionalism maintains merely that we can, *in principle*, correctly identify a given mental state through objective observation, provided we are able to observe *all* behaviour that might be relevant; but of course, in practice, this is not always possible. We have not observed every detail of Jack's behaviour towards Jill, or the prisoner's behaviour towards other people in general, so it is possible, on our limited knowledge, that their statements are misleading. Likewise, in the Turing test, the computer may deceive us, in various ways; but with respect to the questions which the Turing test properly addresses, *this does not matter.*

Imagine that we enter into conversation with another human being, in an Internet chat room. Of course, we are well aware that the person we are speaking to may conceal or

misrepresent his or her beliefs, feelings and emotional attitudes (as indeed, *we* may do ourselves); however, we do not doubt that our interlocutor possesses *some* such human mental attributes. In general, we readily attribute *sentience* to other human beings, without necessarily attributing honesty. Suppose now that, to our complete surprise, it is revealed to us that the person we have been speaking to is, in fact, a computer. Surely, the functionalist will say, we are entitled, in this circumstance, to maintain our conclusion with respect to sentience. On this view, in the case of the *general* features of consciousness and intelligence (as opposed to particulars), functionalism implies that a sufficiently well-stimulated McCoy just is, necessarily, the real McCoy.

Thus, if we accept eliminative, behavioural functionalism, it would seem that we are obliged to reject the lying game argument and accept both Strong AI and the validity of the Turing test. But does this sort of functionalism really rest on solid foundations? Or are there any considerations which might render it doubtful?

One vital implication of the foregoing definition leaps to the eye. If we can establish that there exist mental states, or characteristic features thereof, which *fail* to be captured within a behavioural functional definition, then Turing test functionalism is evidently false—and thus, the supposed justification for strong AI and the validity of the Turing test as a criterion for sentience will collapse.

3 PHENOMENAL QUALITIES

Whenever we reflect upon the subjective character of a sensation or feeling—the blueness of a cloudless sky, the taste of honey, the timbre of a flute, the pain of a toothache, the joy of a kiss, the warmth of summer sunlight on the skin—we confront what is commonly agreed to be one of the great mysteries of conscious experience. Philosophers call these peculiar (yet, very familiar) characteristics *qualia*, or phenomenal qualities. Most are inclined to agree with David Chalmers (1996) when he says that they present any prospective theory of the conscious mind with a 'hard problem'—but some have complained that the standard conception of qualia is loaded with unwarranted, dualistic claims; that it smuggles into our discourse an *a priori* assumption that conscious experience is constituted by strange, ineffable, absolutely irreducible, *intrinsic* properties of—something, we know not what (see, e.g., Dennett 1988). It is important, therefore to define our terms carefully, in order to explicitly avoid any unwarranted, tendentious connotations.

I have suggested, elsewhere, that phenomenal qualities may be usefully defined—without begging any metaphysical questions—as introspectible features of first person experience which we find ourselves utterly unable to describe formally—i.e., purely in terms of structure and/or dynamics (Clifton 2004 [a]). In this deliberately neutral definition, phenomenal qualities are not *postulated* as theoretical entities, endowed with mysterious metaphysical properties, whose existence we propose in an effort to *explain* the nature of conscious experience; rather, they are *pointed out*, as salient features of the way things seem to conscious people. For any theory of the nature of consciousness, a given phenomenal quality is not, in these terms, a hypothetical *explanans*, but rather an evident *explanandum*—and the puzzle it presents to us consists largely in the fact that that it is extraordinarily hard to describe. Thus, particular phenomenal qualities can only be individuated *ostensively*, by specifying the conditions under which we experience them; we cannot, at least with our present knowledge, verbally express or otherwise communicate the nature of their particular subjective characteristics, in the way that, for instance, I can define the notion of a dodecahedron to my eccentrically reclusive neurologist friend, Mary[14] (who has never seen one), so that she can subsequently recognise this particular

[14] See Jackson (1981)

geometric form on sight. I find myself, by contrast, absolutely at a loss to do the same for the subjective character of the colour *blue*.

I consider the foregoing, broad definition to include the valuative properties of first-person experience; that is to say, the vivid impressions we have that our thoughts, sensations and emotional experiences are, for us, in varying degrees, 'good' or 'bad' *in themselves*—regardless of any functional roles which they might also serve in terms of survival or reproduction. It seems just as difficult, for instance, to formally describe our impression of the *intrinsic disvalue*, or *unpleasantness*, of a pain as it is to describe a simple sensation such as the colour *yellow*. Our sense of there being varying degrees of intrinsic value within particular experiences seems essential to our awareness of having such motivational and emotional states as conscious preferences, desires, intentions, hopes, fears, sorrow and joys; furthermore, the presumption that other minds (whether human, animal, alien or artificial), possess, or are capable of possessing, at least roughly similar states seems essential to our regarding them as appropriate subjects of moral concern—and also, perhaps, as having 'rights' that ought to be upheld.

There are two further points that I would like to clarify about the notion of phenomenal qualities—implicit though they may be in the present definition. First, we must be careful not to confuse indescribability of content with ignorance of origin. Phenomenal qualities are not, merely, *intuitions*, *hunches, inspirations, revelations*, or other mental processes in which we come to know or believe something, or generally, arrive at some state of mind, without being able to provide an account of exactly *how* we got there. Here, inexplicably is simply the result of a *lack of information*; we do not have introspective access to the subconscious processes which give rise to our hunches. In the case of a phenomenal quality, the situation is reversed: we are very much aware of a certain kind of qualitative information; and it is this *content* that we are unable to describe. Clearly, if there were *no such content*, but merely subconsciously mediated discriminations, there would be nothing in conscious experience for us to find 'impossible to describe'; on the contrary, they would merely be factual results; inexplicable convictions, whose *origins* we cannot explain.

Second, phenomenal qualities are not *epiphenomena*—at least, not in the strong sense of having no causal powers. It is quite clear that they do, in some sense, play a causally active role—not least, because we are able to talk about them, and comment upon their peculiarity. Moreover, we know that they are closely associated with the occurrence of various perceptual discriminations and emotional states. Their primary functional roles, therefore, appear to be representational and motivational. Each sensory phenomenal quality, for instance, seems to serve as an introspectable token which stands for the occurrence of a particular discrimination. The valuative dimension of qualitative experience, on the other hand, appears to function as a spur to action. Our conscious impressions of the intrinsic *pleasantness* of joys and *awfulness* of pains certainly seem, at least subjectively, to play a significant causal role in the initiation and control of voluntary behaviour.[15]

Notwithstanding the value and importance we ordinarily attribute to phenomenal qualities, and their familiarity within everyday experience, it must be admitted that their fundamental nature is obscure and difficult to explain. Whichever metaphysical theory we prefer with respect to the relation between body and mind, phenomenal qualities seem to present us with a number of remarkably perplexing philosophical and scientific problems.

Any materialist theory, for example, rests upon the basic claim that phenomenal qualities are nothing other than relational physical properties—which is to say that they consist ultimately of

[15] Further support for this common-sense view is provided by evidence of the effects of certain injuries (e.g., lesions in the medial thalamus) and interventions (morphine, frontal lobotomy) on both phenomenology and behaviour. A patient may be fully aware of a normally painful stimulus, yet take no action to avoid it; all of the familiar sensory phenomenal qualities by which it is recognised our present, but the awfulness of it is gone, and the patient remains indifferent (see Hardcastle, 1999).

structural and/or dynamic relationships between instances or combinations of fundamental physical entities.[16] On this view, the fact that we are unable to describe phenomenal qualities in such terms can only be explained by the suggestion that, *qua* physical phenomena, they are extraordinarily complex and subtle in nature; possessed of various *as-yet-unknown* relational properties which, give rise, *somehow*, to their familiar qualitative characteristics—whilst also, *somehow*, utterly obscuring all evidence of their true, underlying, structural/dynamic nature. A remarkable proposition, indeed.[17]

Various forms of mentalism, on the other hand (e.g., idealism, panpsychism, substance dualism) suggest that the nature of, at least, the simplest of phenomenal qualities should be accepted as *brute fact*, as opposed to a complex puzzle in need of explanation. On this view, phenomenal qualities are not relational properties of any kind. We may interpret them, instead (for example), as intrinsic properties of an inherently self-aware mental substance—and think of these simple properties as the fundamental constituents of conscious experience itself. This would mean, however, that phenomenal qualities must, *somehow*, arise in 'mindspace' in response to physical events in brain, by means of mysterious, *yet-to-be-discovered* psycho-physical laws. Furthermore, if we are to avoid the paradoxes of epiphenomenalism, we must assume these laws to be interactive; thus, the non-physical mind must *somehow* manipulate the course of events in the brain—in some way that would surely require a revision of the laws of physics.[18]

Whatever they may be, it is evident that for many philosophers, phenomenal qualities are extremely *embarrassing*—and not only because the alternative metaphysical theories of their nature seem (almost) *equally* outlandish and suffused with mystery. In addition, it is clear that at least on our present knowledge, their qualitative characteristics cannot be conveyed in a conventional functional definition.

Suppose that we define a phenomenal quality, in the prescribed behavioural functionalist manner, as a persistent disposition—when asked certain phenomenological questions—to give answers of a certain sort, such as: "…when I look up at the sky (when the weather is fine and the sun is shining), I experience a qualitative sensation, which I refer to as *blue*, but I cannot begin to describe it, or explain the way it seems." Clearly, such a definition utterly fails to capture the subjective character of *blue* colour sensations. If we presented it to someone who, blind from birth, has never experienced colours, it surely wouldn't help them to *recognise* particular sensory qualities—when enabled, through surgery, to experience them the first time.

It might be speculated that functional descriptions of phenomenal qualities—unavailable on our present, limited knowledge—may be possible *in principle*.[19] There is, of course, no evidence whatsoever to suggest that this is so; but even if we take the hypothesis seriously, it is clear that it offers no support to Turing test functionalism. There is no obvious reason why such a functional analysis, if achievable at all, should be possible in terms of *verbally expressible* behavioural dispositions alone; it might, perhaps, require a detailed description of internal functional process—or of non-verbal behaviour patterns which are highly specific to human physiology (e.g., blushing, pupil dilation, body-language etc). Even if we stipulate that our 'yet-to-be-discovered functionalism' *must* be of the 'verbally-expressible-behaviour' kind, it remains distinct from conventional Turing test functionalism; for until we have gone beyond speculation and *solved* the mind-body problem in this fashion, interrogators in the

---

[16] Assuming that fundamental physical entities do not, in themselves, possess any mental characteristics

[17] In Clifton (2004) [a] I explore the implications of this 'cryptic complexity hypothesis' and the difficulties it raises for materialism.

[18] See Clifton (2004) [b] for a critical discussion of epiphenomenalism and various alternative dualist views.

[19] This view is popular with materialist philosophers; see e.g. (Dennett, 1991 ch. 12). My 'description argument' (Clifton, 2004 [a]) suggests that it is highly implausible—but does not categorically excluded it.

Turing test will not know what subtle verbal cues to look out for, in order to confirm the genuine occurrence, in the computer, of qualitative mental states.

Hence, whatever the fundamental nature of phenomenal qualities may be, it is quite clear that in this case at least, eliminative, Turing test functionalism is false. With respect to the question of artificial sentience, it is possible, therefore, to infer some useful information from their mere existence—without first having to determine exactly what they *are*, or definitely commit ourselves to either side of the metaphysical pale.

The only option that remains available to the Turing test functionalist is to assume what I like to call the 'Faith-healer of Deal position'[20]—and flatly *deny* the existence of phenomenal qualities. On this radically eliminative view, the allegedly introspectible, qualitative content which people *claim* they are familiar with (and are unable to describe), *simply isn't there*; our *belief* in its existence is utterly unfounded.[21] It seems to me, however, that this remarkable thesis provides not so much a defence as a *reductio ad absurdam* of third person absolutism. To mention but one rather obvious weakness, it begs the question: what is it like to *experience the illusion* that one is aware of a phenomenal quality? Surely, for such a hallucination to be persuasive, it must be exactly like experiencing some special kind of conscious content, which one cannot formally describe. In order to mislead us, therefore, the hypothetical illusion must itself be an instance of the very phenomenon which the eliminativist seeks to deny.[22]

In any case, it is one thing to question (coherently or otherwise) the reliability of our judgments as to the way things subjectively seem—and quite another to demonstrate that an entire class of such judgments are definitely and invariably false—and that their objects, the qualitative 'seemings' which *constitute* our experience, do not exist at all. Even if the eliminativist were able—*per impossible*—to refute the foregoing *reductio* and produce a coherent and plausible account of how the 'illusion hypothesis' *might* work, they would also need—in order to vindicate Turing test functionalism—to provide positive *proof* that such a mechanism *actually occurs*. Of course, such arguments and evidence are as conspicuously absent within the philosophical and scientific literature as phenomenal qualities are abundantly present, in the first-person experience of *most* normal human beings. Unless and until these circumstances dramatically change, we may surely remain confident that eliminative functionalism is false.

4 FUNCTIONAL REDUNDANCY AND THE LYING GAME

Consider once again the functional roles which phenomenal qualities appear to fulfil. We become aware of a particular kind of qualitative experience whenever we make some sort of sensory discrimination, or enter into a particular emotional state; hence, phenomenal qualities seem to serve, at least, as representative and/or motivational tokens which signal the occurrence of functionally definable brain states. However, the particular kind of subjective content which characterises each phenomenal quality does not appear to be required by these functional roles; indeed, qualitative content *of any sort* does not seem necessary—for if the objective, behavioural functionality of a mental state can be fully defined without describing the subjective character of the accompanying phenomenal quality, then surely, this functionality can be implemented without it. In other words, phenomenal qualities seem to be *functionally redundant*—surplus to purely practical requirements, rather like the whimsical, decorative embellishments we sometimes see on pieces of antique machinery. Since the functionality of the machine is *multiply realisable*, a new, starkly efficient, modernist sort of

---

[20] There was a faith-healer of Deal / Who said, "Although pain isn't real, / when I sit on a pin / and it punctures my skin, / I dislike what I fancy I feel." Anon.

[21] For various arguments which have been marshalled to this effect, see e.g. Wittgenstein (1953), Sellars, (1956), Feyerabend (1963a), (1963b), Rorty (1965, 1970, 1979), Dennett (1988, 1991), P. M. Churchland (1981, 1984), P. S. Churchland (1986), For a critique of the eliminative view, see Clifton (2004) [a] §4.

[22] See Clifton (2004) [a].

machine can easily be created, without any such decorations—whose functional output is indistinguishable from that of the original antique. Similarly, in place of qualitative tokens, it seems perfectly plausible, *in principle*, that our brains could be re-wired so as to use a simpler, less metaphysically mysterious signalling system—such as binary code-numbers, instantiated in patterns of neurological activity. The electro-chemical machinery of the brain would otherwise work just the same—except, of course, that we would no longer claim to be acquainted with phenomenal qualities.

In the light of these considerations, it might be supposed that a machine which possesses a human-like system of qualitative mental-state coding would be readily distinguishable, in a Turing test, from one which entirely lacks such qualitative experience. All we would need to do, as interrogators, is to ask a number of phenomenological questions. Should the computer provide typically human answers we might judge it to be sentient; but if it denies all knowledge or awareness of indescribable subjective qualities, or of sensations which seem unanalysably 'good' or 'bad' in themselves, then we may reasonably assume that it is non-sentient, and does not possess such mental attributes. However, this procedure requires us to assume that the computer will *tell the truth* about its mental states—an assumption which is not permissible under the standard conditions of the Turing test.

Consider now the functionalist reply to the 'lying game' argument: that a falsified statement about private experience is evidence for a behavioural disposition *of some sort*; that it is, for *some* kind of mental attribute—albeit not the *particular* feeling, belief or sensation that it purports to express. This standard defence of the Turing test depends upon the assumption that all mental attributes are fully definable in terms of observable, functional criteria—and hence, that they cannot possibly be separated from their functional role. In the case of phenomenal qualities, as we have seen, this assumption seems to be false. It would be trivially easy to devise a computer program—entirely lacking in qualitative experience—which nevertheless passes the phenomenological Turing test, by *cheating*. All one would have to do is program the computer to respond to such questions with typically human answers, such as: "…when I look up at the sky, I experience a qualitative sensation, which I refer to as *blue*, but I cannot begin to describe it, or explain the way it seems." A disposition to claim awareness of phenomenal qualities—and also to perform the basic sensory discriminations or goal-directed behaviours associated with them—is, in functional terms, *multiply realisable*; but since phenomenal qualities are functionally redundant, not all such possible realisations need involve them, as part of the mechanism. As interrogators—*without access to the relevant source code*—we cannot take the mere fact that a computer makes human-like statements about its 'phenomenal qualities' as evidence that it *actually possesses them*—any more than we can safely rely upon its claims to have long blonde hair, a diploma in modern dance and a fabulously wealthy uncle in Pasadena.

Thus, when we take into account the functionally redundancy of phenomenal qualities, the 'lying game' argument defeats Turing test functionalism. We may summarise this argument as follows:

(1). Conscious experience is characterised by various kinds of introspectible, *qualitative* content which we find ourselves unable to describe or express in abstract, formal terms.
(2). There is no coherent or credible eliminative argument against the *existence* of such 'phenomenal qualities', *in this straightforward, metaphysically neutral sense*.
(3). The existence of such qualitative content is incompatible with eliminative, behavioural functionalism: *on our present knowledge at least*, the nature of qualitative experience cannot be captured in any such functional definition of a mental state.
(4). It follows that as far as we can tell, phenomenal qualities are *functionally redundant*; that is, they are separable, in principle, from the representational and motivational functional roles which they appeared to serve.

(5). It is possible, therefore, for a machine to reproduce all of the behaviour we associate with these functional roles, in the *absence* of phenomenal qualities. It is also possible for such a machine to *lie* about its mental states, claiming acquaintance with "indescribable" phenomenal qualities—which in fact, it does not possess.
(6). Therefore, at least on our present knowledge, consistent success in passing the Turing test fails to establish whether or not a given machine is sentient.

While the foregoing lying game argument invalidates the Turing test criterion for sentience, it falls short of disproving the other major assumptions of Strong AI; most importantly, the claim that a sufficiently powerful and suitably programmed computer will be sentient. Our next task, must be to consider two questions which hitherto remain unanswered: is a sentient machine possible *in principle*—and if so, is it *likely* that our current research programmes in artificial intelligence will ultimately give rise to such a device?

5 SENTIENT MACHINES

We know that sentient machines are possible, in principle—because *we*, at least in some sense, are sentient machines. This claim is not, I think, quite so contentious as it might seem at first sight. Whether the various phenomenal qualities with which I am familiar are (as materialists claim) somehow instantiated physically in neurological activity, or (as mentalists would suggest) are elicited in my non-physical *psyche* by virtue of psycho-physical laws, it seems clear that their presence within conscious experience is contingent (at least, some of the time) upon the occurrence of certain *physically determined conditions* within my brain.

Of course, other human beings have brains, very similar to my own—and it seems highly unlikely that mine is unique, or untypical, in its ability to satisfy these conditions; thus, the lying game argument against Turing test functionalism does not lend any credence to agnosticism or solipsism with respect to other *human* minds. These sceptical hypotheses imply not only that all/some other human brains fail, for some unknown reason, to meet the requisite conditions for qualitative experience, but also that such 'zombie-brains' are, for some further peculiar reason, *programmed to lie*—deceptively answering phenomenological questions in the same sort of way that I do. Such a bizarre, self-censoring 'coincidence of anomalies' is, of course, extremely improbable. The *combination* of biological and hetero-phenomenological similarity therefore provides, I think, sufficient justification for the attribution, by sentient human beings, of qualitative experience to other human beings; likewise, biological and behavioural similarities provide excellent grounds for attributing sentience—i.e., qualitative experience, emotion, pleasure and pain—to (at least) the higher animals (see e.g., Masson & McCarthy, 1995).

Notwithstanding these arguments based on biological similarity, we cannot safely assume, on our present knowledge, that the special physical criteria required for qualitative experience may *only* be fulfilled by biologically mediated, neural events. It may be possible, for all we know, that a machine or artefact whose physical constitution is considerably different from ours might nevertheless satisfy the conditions in question. However, until we are able to firmly identify the requisite conditions, we likewise have no reason to presume that any given machine, designed simply to generate intelligent *behaviour*, will happen to fulfil them. Indeed, there are ample grounds for suspicion that such an eventuality is highly improbable.

If materialism is true, then phenomenal qualities must be instantiated in complex patterns of neurological activity—far greater complexity, we must surely assume, that the minimum required to reliably individuate the discriminations and functional states with which they are associated.[23] It follows that this deeply mysterious 'qualitative complexity' must constitute a

[23] As argued above, the minimum functional requirement—even allowing plentiful redundancy to minimise errors—would be a simple binary code number.

remarkably inefficient use of bandwidth—that is, the overall data-handling capacity of an information processing system. If, however, we continue to design our computers and our artificial intelligence programmes for maximum information-processing performance, we will naturally seek to avoid unnecessary representational complexity of any kind. Suppose, on the other hand, that we develop artificial intelligence using some trial-and-error evolutionary process, in which programs are able to *mutate* in some uncontrolled way, and we then select those which demonstrate the most objectively advanced cognitive abilities. It still seems highly unlikely that a mutation which led to inefficient use of bandwidth would survive the selection process—or that random mutations would just happen to give rise to the kind of peculiar complexity which instantiates phenomenal qualities in our own case.

Materialists might speculate that phenomenal qualities serve some unknown function, in addition to their role as representative tokens; and that this mysterious property contributes to our information-processing abilities in some way which has definite advantages over a simpler, non-qualitative, representative system. However, we have no means of knowing whether or not this speculation is correct, not any reason to suppose that this mysterious functionality is either necessary for truly intelligent behaviour, or likely to arise in *any* system which develops intelligence through an evolutionary process of mutation and selection. Its evolution, in our own case, might have arisen due to some uniquely improbable combination of mutations, or some special set of highly unusual environmental circumstances which happened to obtain at some distant point in our evolutionary past. Nor, on our present knowledge, could we 'assist' the evolutionary process; for until we achieve a comprehensive materialistic solution to the mind-body problem, we have no way of knowing what sort of complexity is required to instantiate phenomenal qualities.

If, on the other hand, some form mentalism is true, then qualitative experience occurs in the mind in response to physical events in the brain, in accordance with psycho-physical laws. A number of theorists have proposed possible ways in which such laws might work, involving, for example, peculiar quantum-events in the brain at the synaptic or sub-neuronal level.[24] Until these bold speculations are confirmed and elucidated, however, we have no reason to suppose that a digital computer—whose basic components are strikingly dissimilar to neurons in their constitution and operation—will fulfil the conditions which these hypothetical psycho-physical laws require to be satisfied.

For the reasons I have outlined here, then, it is highly unlikely that any *given* information-processing device, designed to generate intelligent behaviour, or to *simulate* sentient behaviour, will just happen to satisfy the yet-to-be-identified and (as far as we can tell) *functionally redundant* physical conditions—obtaining, in our own case, within human brains—which are necessary, or sufficient, for the occurrence of qualitative conscious experience. On the other hand, we cannot altogether exclude the possibility that such a machine would be sentient—and the Turing test is unable to settle the matter.

6 COMPUTING MACHINERY AND SENTIENCE

In §4 we considered the possibility of a specialised Turing test, in which the interrogator is required to ask phenomenological questions and we attribute sentience to those machines which pass the test, by providing typically human answers. Given the present definition of phenomenal qualities, such questions might include:

(A) Are you aware, in your own experience, of phenomenal qualities?
(B) Could you formally describe the subjective content of a colour sensation?
(C) Do some of your experiences seem intrinsically desirable, or good-in-themselves?

[24] For arguments in support of this view, see e.g. Averill & Keating (1981), Beloff (1994a, 1994b) Eccles (1986), Beck & Eccles (1992), Larmer (1986), Libet (1994), Lowe (1992, 1993).

(D) Do other experiences seem intrinsically unpleasant, or bad-in-themselves?
(E) Can you describe or explain the nature of this sense of subjective value?

We would expect most normal human beings to say "yes" to (A), (C) and (D) and "no" to (B) and (E). Turing test functionalism implies, therefore, that we should attribute sentience to any machine which consistently gives the same answers. Our objection to this approach, of course, was that a non-sentient machine could easily be made to respond in this way; not because it actually *experiences* qualitative mental states, but rather because it has been programmed to *lie*. On the other hand, if we were to demand access to the source-code, in order to determine whether or not our suspicions are justified, we would undermine Turing's strictly agnostic, behaviourist approach—which was introduced in order to decisively refute any *a priori* assumption that machines cannot think. If we accept the need for this restriction, we are faced with an inescapable dilemma: and there is no way tell whether or not the machine is sentient.

Fortunately, it seems clear that we can safely dispense with the blind methodology. As I argued in §5, we can defeat the anti-mechanistic prejudice by means of reason alone: an artificial, sentient machine is possible in principle, for insofar as conscious experience is causally dependent upon neuro-chemical events in the brain, *we are, at least in that sense, sentient machines*. There is no obvious reason why the physical conditions sufficient for the generation of consciousness in the brain cannot be satisfied by some artificial mechanism. Hence, if we had a good reason to believe that a given machine is neither programmed nor trained to lie about its mental states, then it would seem reasonable to take its consistently expressed phenomenological claims at face value.

A straightforward solution to our empirical predicament now suggests itself. We may replace Turing's Imitation Game with an *Introspection Game*, in which the blind condition is removed and a new rule is introduced, allowing investigators to closely examine the internal processes associated with the machine's generation of replies to phenomenological questions, so as to establish whether or not these responses are nothing more than pre-programmed or imitative lies. Let us assume, without loss of generality, that the machine to be evaluated is a digital computer of a more-or-less conventional kind. In this case, we may call the new rule the '*open source, open mechanism*' condition. The investigators must be provided with complete, open-source access to the software—and also unrestricted freedom to examine and monitor the functioning of the hardware, in real time, while the computer responds to phenomenological questions.

Let us now consider the possible outcomes which might arise from such an investigation.

(1) The machine reports no awareness of phenomenal qualities.
(2) The machine claims awareness of phenomenal qualities, but analysis of the source-code and of the machine's functional operation reveal that its phenomenological reports are merely pre-programmed (or learned) imitative responses.

In these circumstances, we may conclude that the machine is (almost certainly) non-sentient. As we have seen, it seems overwhelmingly likely that machines constructed on principles that we currently understand would produce one of these two results; but it nevertheless remains possible that that the outcome of our investigation may completely surprise us:

(3) The machine claims awareness of phenomenal qualities, but the investigators can find no mechanism for, or evidence of pre-programmed or learned mendacity.

In this situation, we may conclude, at least tentatively, that the machine is probably sentient. Our theoretical interpretation of this result (and our confidence therein) will be influenced by further details of the investigators' report, which might reasonably be expected to announce one of the following findings:

(4) the AI program represents discriminative information states in some strangely complex way, such that it is able to recognise this qualitative content but not describe it.
(5) discriminative information states are represented straightforwardly within the system. However, phenomenological reports (and other conscious and involuntary behavioural responses) arise in a way which consistently violates either the logic of the program or the known laws of physics, or both.

The first of these alternatives would be consistent with a materialist account of sentience, while the second is suggestive of interactionist dualism. In either case, our original conclusion that the machine is sentient—together with the appropriate metaphysical interpretation—would be greatly reinforced if comparisons between the investigators' discoveries and studies of the brain are able to establish a parallel. Independently of the Introspection Game investigation, neuroscientists might have already discovered that either:

(6) there exist strange and profoundly complex patterns of neural activity in the brain which are associated with experiences of phenomenal qualities
(7) peculiar physical anomalies occur frequently in the brain, and seem to be associated with experiences of phenomenal qualities and the utterance of phenomenological statements.

In other words, it is conceivable that similarities will be observed between either complex patterns, or physical anomalies in both brain and machine.[25] Such a match, in either case, would strongly support our original conclusion that the machine is sentient, in a human-like way. In addition, such findings would seem to justify a metaphysical interpretation—in one of two possible directions. A match between (4) and (6) would provide a powerful confirmation of the materialist theory of mind, whereas (5) and (7) would serve to vindicate the mentalist view. In the latter eventuality, we might reasonably conclude not only that the Tin Man has a 'heart', but also that—in our own case as well as his—there really is a ghost in the machine.

References

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition* (DSM-IV), Washington DC: American Psychiatric Press

Averill, E. W. & Keating, B. (1981) 'Does interactionism violate a law of classical physics?' *Mind* 90:102-7.

Beck, F. and Eccles, J. C. (1992) "Quantum aspects of brain activity and the role of consciousness." *Proceedings of the National Academy of Science*, 89(23): 11357–11361.

Beloff, J. (1994a) "Minds and machines: a radical dualist perspective." *Journal of Consciousness Studies* Vol 1 No. 1 pp. 32–37

Beloff, J. (1994b) "The mind brain problem." *Journal of Scientific Exploration*, Vol 8 No 4.

Block, N. (1979) 'Troubles with Functionalism.' In P. French, T. Uehling, and H. Wettstein *eds*. *Perception and Cognition*. Minneapolis: University of Minnesota Press

Block, N. (1981) *Psychologism and behaviorism.* Philosophical Review 90:5-43

Boden, M., ed. (1980) *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

Bridgman, P. W. (1927) *The Logic of Modern Physics*. New York: Macmillan

Brink, D. O. (1989) *Moral Realism and the Foundations of Ethics*. Cambridge University Press.

Broderick, D. (1999) *The Last Mortal Generation*. Sydney: New Holland/Striek

Broderick, D. (2001) *The Spike: How Our Lives are being Transformed by Rapidly Advancing Technologies.* New York: Forge.

[25] It seems highly unlikely that a mismatch would occur, e.g. (4) and (7) or (5) and (6). I leave the reader to ponder how such a peculiar outcome should be interpreted.

Brooks, R. A. (2003) *Flesh and Machines*. Vintage Books
Chalmers, D. (1996a) *The Conscious Mind.* Oxford: Oxford University Press.
Chalmers, D. J. (1996b) 'Minds, machines, and mathematics.' *Psyche* 2:11-20.
Chomsky, N. (1959) 'Review of *Verbal Behavior* by B. F. Skinner.' *Language* 35: 26-58.
Churchland, P. M. (1981). 'Eliminative Materialism and the Propositional Attitudes.' *Journal of Philosophy* 78: 67-90.
Churchland, P. M. (1984) *Matter and Consciousness*. Cambridge, MA.: MIT Press
Churchland P. M. (1996) *The Engine of Reason, The Seat of the Soul: A Philosophical Journey into the Brain.* MIT Press.
Churchland P. M. and Churchland P. S. (1998) *On the Contrary: Critical Essays, 1987-1997.* MIT Press
Churchland P. S. (1986) *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* MIT Press
Churchland P. S. (2002) *Brain-Wise: Studies in Neurophilosophy.* MIT Press
Clark, A. (1997) *Being there: putting brain, body and world together again*. MIT Press
Clark, A. (2000) *Mindware: An introduction to the philosophy of cognitive science.* Oxford University Press.
Clifton, A. (2004) [a] An empirical case against materialism. Unpublished MS.
Clifton, A. (2004) [b] *Res cogitans*. Unpublished MS.
Clifton, A. (2004) [c] The hazards of silicon heaven. Unpublished MS.
Cotterill, R. (1989). *No Ghost in the Machine: Modern Science and the Brain, the Mind and the Soul.* London: Heinemann.
Cotterill, R. (1997). On the mechanism of consciousness. *Journal of Consciousness Studies*, 4(3), 231-247.
Dennett, D. (1987) *The Intentional Stance*. MIT Press.
Dennett, D (1988) *Quining Qualia.* in A. Marcel and E. Bisiach, eds, *Consciousness in Modern Science*, Oxford University Press 1988.
Dennett (1997) 'Consciousness in human and robot minds', In Ito, Masao, Miyashita, Yasushi and Rolls, Edmund T. *Cognition, Computation, and Consciousness.* Oxford University Press
Dery, M (1996) *Escape Velocity: Cyberculture at the End of the Century.* Grove Press.
Eccles, J. C. (1986) "Do mental events cause neural events analogously to the probability field of quantum mechanics?" *Proceedings of the Royal Society of London* B227: 411–28.
Elton, M (1997) ''Robots and Rights: The Ethical Demands of Artificial Agents', *Ends and Means, Journal of the Aberdeen Centre for Philosophy, Technology and Society*, New Series, 2: 19-23
Elton M (2000) 'Should Vegetarians Play Video Games?', *Philosophical Papers* 29(1) 21-42
Feyerabend, P. (1963a) 'Mental Events and the Brain' *Journal of Philosophy* 60: 295-296.
Feyerabend, P. (1963b) 'Materialism and the Mind Body/Problem' *The Review of Metaphysics* 17:49-66.
Fodor, J. (1975) *The Language of Thought*. New York: Crowell.
Fodor, J. (1980) 'Methodological solipsism considered as a research strategy in cognitive science.' *Behavioral and Brain Sciences* 3:63-73.
Fodor, J. (1981) 'The Mind Body Problem.' *Scientific American* 244:1:114-123
Fodor, J. (1990) *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
Friedman, M (1999) *Reconsidering Logical Positivism* Cambridge University Press.
Furse, E (1995) 'A theology of robots', *unpublished ms.* available online at: *www.comp.glam.ac.uk/pages/staff/efurse/Theology-of-Robots/A-Theology-of-Robots.html*
Grand, S (2001) *Creation: Life and How to Make It*. Harvard University Press
Gregory R. (1987) 'In defence of AI: a reply to Searle.' In C. Blakemore and S. Greenfield *Eds.,* (1987) *Mindwaves.* Blackwell.

Grush, R. & Churchland, P. (1995) 'Gaps in Penrose's toiling.' In T. Metzinger, ed. *Conscious Experience*. Ferdinand Schoningh.
Hadley, R. F. (1987). Godel, Lucas, and mechanical models of mind. *Computational Intelligence* 3:57-63.
Harman, G (1977) *The Nature of Morality*. Oxford: Oxford University Press
Harman, G (1999) 'Wide Functionalism' in: *Reasoning, Meaning, and Mind.* Clarendon Press.
Hofstadter, D. R. (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books

Hofstadter, D. R. and Dennett, D. C. *Eds*. (1981) *The Mind's I.* Bantam Books
Holst, G. (2001) 'Should Robots be Slaves? Is it too late to consider robot rights?' *unpublished ms.* available online at: *www.sfu.ca/~gholst/RobotSlaves/robotslaves.html*
Jackson, F. (1982) 'Epiphenomenal qualia.' *Philosophical Quarterly* 32:127-136
Johnson-Laird, P (1987) 'How could consciousness arise from computations in the brain.' In C. Blakemore and S. Greenfield *Eds.,* (1987) *Mindwaves*. Blackwell.
Johnson-Laird, P (1988) The *Computer and the Mind.* Fontana
Kurzweill, (1999) *The Age of Spiritual Machines.* Penguin Books.
Lewis, D. (1980) 'Mad Pain and Martian Pain.' in N. Block, ed., *Readings in the Philosophy of Psychology, Vol. I.* Harvard University Press
Libet, B. (1994) A testable theory of mind-brain interaction. *Journal of Consciousness Studies* 1:119-26.
Lowe, E. J. (1992) The problem of psychophysical causation. *Australasian Journal of Philosophy* 70:263-76.
Lowe, E. J. (1993) The causal autonomy of the mental. *Mind* 102:629-44.
Lucas, J. R. (1996) 'Mind, machines and Gödel' *Philosophy*, 36, 112-27
Lucas, J. R. (1996) 'Mind, machines and Gödel: A retrospect.' In P. Millican & A. Clark, eds. *Machines and Thought.* Oxford University Press.
Lycan, W. (1987) *Consciousness* Cambridge MIT Press
Lycan, W. (1996) *Consciousness and Experience*. MIT Press
Mackie, J. L. (1977) *Ethics: Inventing Right and Wrong*. Harmondsworth, Penguin Books
Masson, J. M. and McCarthy, S. (1996) *When Elephants Weep: The Emotional Lives of Animals.* Dell Publishing Company.
McCarthy (1995) 'What has AI in Common with Philosophy?' *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 2 pp.* 2041-2044
McDermott, D. (2001) *Mind and Mechanism*. MIT Press.
McNally, P., & Inayatullah, S. (1988). 'Rights of Robots.' *Futures*, 20(2), 119-136.
Misak, C. J. (1995) *Verificationism: Its History and Prospects*. Routledge
Moor, J.H. (1976) 'An Analysis of the Turing test', *Philosophical Studies* 30, 249–257.
Moor, J.H. (1987) 'Turing Test' in S.C. Shapiro, ed., *Encyclopedia of Artificial Intelligence*, New York: John Wiley and Sons, pp. 1126–1130.
Moor, J.H. (2000), 'Turing Test', in A. Ralston, E.D. Reilly, D. Hemmendinger, eds., *Encyclopedia of Computer Science*, 4th edition, London: Nature Publishing Group, pp. 1801–1802.
Moor, J.H. (2001) 'The Status and Future of the Turing Test' *Minds and Machines* 11: 77–93
Moravec, H. (1990) *Mind Children: The Future of Robot and Human Intelligence.* Harvard University Press.
Moravec, H. (2000) *Robot: Mere Machine to Transcendent Mind.* Oxford University Press
Murphy, J. (1990) *Pragmatism from Pierce to Davidson.* Westview Press.
Paul, G. S. and Cox, E. (1996) *Beyond Humanity: Cyberevolution and Future Minds*. Charles River Media
Penrose, R. (1989) *The Emperor's New Mind*. Oxford: Oxford University Press

Penrose, R. (1994) *Shadows of the Mind*. Oxford: Oxford University Press
Pierce, C. (1878) 'How to Make Our Ideas Clear' *Popular Science Monthly* 12: 286-302.
Pojman, L. P. (1994) *Ethics: Discovering Right and Wrong*. (2nd edition) Wadsworth Publishing.
Polanyi, M. (1958) *Personal Knowledge: Towards a Post-Critical Philosophy.* Chicago: University of Chicago Press.
Polanyi, M. (1966) *The Tacit Dimension.* Doubleday, New York, 1966
Preston, J and Bishop, M. *Eds* (2002) *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence.* Oxford University Press
Putnam, H. (1962), "What Theories are Not", in E. Nagel, P. Suppes and A. Tarski (eds.), Logic, Methodology and Philosophy of Science, Stanford, Stanford University Press
Putnam, Hilary, (1975), 'Explanation and Reference.' in Mind, Language, and Reality. Philosophical Papers, Vol. II. Cambridge University Press
Putnam, Hilary, (1975), 'The Meaning of Meaning.' in Mind, Language, and Reality. Philosophical Papers, Vol. II. Cambridge University Press
Quine (1951) 'Two Dogmas of Empiricism' *Philosophical Review*, 60(1): 20-43
Rey, G. (1980) 'Functionalism and the Emotions.' in A. Rorty, ed., *Explaining Emotions* Berkeley: Univ. of California Press.
Rey, G. (1996) 'Towards a Projectivist Account of Conscious Experience.' in T. Metzinger, *ed.*, *Conscious Experience.* Paderhorn: Ferdinand-Schoningh
Rey, G. (1997a) Contemporary Philosophy of Mind: A Contentiously Classical Approach. Blackwell.
Rey, G. (1997b) 'A Question About Consciousness.' in, ed. by N. Block, O. Flanagan, and G. Guzeldere, eds., *The Nature of Consciousness.* MIT Press.
Rorty, R (1965) 'Mind-body Identity, Privacy and Categories.' *The Review of Metaphysics* 19:24-54
Rorty, R (1970) 'In Defense of Eliminative Materialism.' *The Review of Metaphysics* 24:112-121
Rorty R. (1972) 'Functionalism, Machines, and Incorrigibility' *Journal of Philosophy*, 69: 203-220
Rorty, R (1979) *Philosophy and the Mirror of Nature*. Princeton University Press.
Ryle, Gilbert. (1949) *The Concept of Mind*. Chicago: The University of Chicago Press
Reichenbach H. (1953) 'The Verifiability Theory of Meaning' in H. Feigl and M. Brodbeck eds., *Readings in the Philosophy of Science*. Appleton-Century-Crofts
Sayre McCord, G., ed. (1988) *Essays on Moral Realism.* Cornell University Press
Schlick M (1959) 'The Turning Point in Philosophy' in Ayer A. J. (ed.) *Logical Positivism*. Free Press.
Searle, J. (1980) 'Minds, brains and Programs.' *Behavioural and Brain Sciences* 3: 417-24
Searle, J. (1984) *Minds, Brains and Science*. Harvard University Press
Searle, J. (1992) *Rediscovering the Mind*. Harvard University Press
Searle, J. (1997) *The Mystery of Consciousness.* New York, New York Review Press
Sellars, W. (1956) Empiricism and the philosophy of mind. In H. Feigl and M Scriven, eds., *The Foundations of Science and Concept of Psychology and Psychoanalysis*. University of Minnesota Press
Shook, J. ed., (2000) *The Chicago School of Pragmatism*. Thoemmes Press. 4 Volumes.
Skinner, B. F. (1953) *Science and Human Behavior*. New York: Macmillan
Sloman, A. (1978) *The Computer Revolution In Philosophy: Philosophy, science and models of mind.* Harvester Press.
Sloman, A (2000) 'Architectural Requirements for Human-like Agents Both Natural and Artificial (What sorts of machines can love?)' In: K. Dautenhahn, Ed., *Human Cognition And Social Agent Technology*. John Benjamins Publishing

Sloman, A and Croucher, M (1981) 'Why robots will have emotions.' *Proceedings IJCAI 1981,* Vancouver.
Skinner, B. F. (1953) *Science and Human Behavior*. New York: Macmillan.
Stevenson, C. L. (1937) 'The emotive meaning of ethical terms.' *Mind*, 46, 14-31.
Stevenson, C. L. (1944). *Ethics and language*. Hew Haven, CN: Yale University Press
Taylor J. (1995) 'Towards the Ultimate Intelligent Machine.' Presidential Address, *World Congress on Neural Networks*, Washington DC, July 17-21.
Taylor J. (1999) *The Race for Consciousness*. MIT Press
Tipler, F. J (1995) *The Physics of Immortality: Modern Cosmology, God and the Resurrection of the Dead.* London: Macmillan
Turing, A. (1950) Computing machinery and intelligence. *Mind* 59: 443-460.
Vinge, V. (1993) 'The Coming Technological Singularity: How to Survive in the Post-Human Era', *Whole Earth Review*, No. 81 pp. 88ff.
Watson, J. B. (1912) Psychology as the behaviorist views it. *Psychological Review* 20: 158-177.
Williams, B. (1973) 'Morality and the Emotions.' In: *Problems of the Self*. Cambridge University Press.
Williams, B. (1982) *Moral Luck : Philosophical Papers 1973-1980.* Cambridge University Press
Williams, B. (1986) *Ethics and the Limits of Philosophy.* Harvard University Press
Wittgenstein, L. (1953) *Philosophical Investigations*. Trans. G. E. M. Anscombe. New York: Macmillan
Yudkowsky E. (2001) 'The Singularitarian Principles (Extended Version' *sysopmind.com/sing/principles.ext.html*
Yudkowsky, Eliezer S. (2003) *Levels Of Organization In General Intelligence*, in Ben Goertzel and Cassio Pennachin, eds. *Real AI: New Approaches to Artificial General Intelligence*
Zuriff, G. E. (1985) *Behaviorism: A Conceptual Reconstruction*. Columbia University Press