Blind man's bluff and the Turing test

Andrew Clifton

Abstract

It seems plausible that under the conditions of the Turing test, congenitally blind people could nevertheless, with sufficient preparation, successfully represent themselves to remotely located interrogators as sighted. Having never experienced normal visual sensations, the successful blind player can prevail in this test only by playing a 'lying game'—imitating the phenomenological claims of sighted people, in the absence of the qualitative visual experiences to which such statements purportedly refer. This suggests that a computer or robot might pass the Turing test in the same way, in the absence not only of visual experience, but qualitative consciousness in general. Hence, the standard Turing test does not provide a valid criterion for the presence of consciousness. A 'sensorimetric' version of the Turing test fares no better, for the apparent correlations we observe between cognitive functions and qualitative conscious experiences seems to be contingent, not necessary. We must therefore define consciousness not in terms of its causes and effects, but rather, in terms of the distinctive properties of its content, such as its possession of qualitative character and apparent intrinsic value—the property which confers upon consciousness its moral significance. As a means of determining whether or nor a machine is conscious, in this sense, an alternative to the standard Turing test is proposed.

1. Cutting the Turing test down to size

I propose to consider the question, *does the Turing test provide a valid criterion for the presence of consciousness?* This should begin, at least, with a definition of the term 'Turing test' (and I promise to return to the more difficult 'consciousness' before the end of this paper). Alan Turing's celebrated 'imitation game' test—in which a computer is programmed to pass itself off as a human being—was originally devised as a means of addressing the question, 'can machines think.' (Turing, 1950). In the standard version of the game, one player, a (human) 'referee', uses a communications terminal of some kind to exchange typewritten messages with an remotely located opponent—and seeks to determine whether this entity is human, or rather, a computer designed to simulate a human personality. If, over a number of trials, a succession of expert referees are consistently unable to make this distinction, the computer is deemed to have won the game—and passed the test. With respect to Turing's initial question, however, there are at least two distinct affirmative interpretations which might be placed upon such a result:

(1) Yes, the machine 'thinks'—at least, in the limited sense that it can match the formidable information-processing abilities which allow humans to *seem human* to other humans, under the restricted conditions of the Turing test.

(2) Yes, the machine *thinks*—in the strong sense of being fully sentient and conscious; that is, possessed of sensations, feelings, beliefs, hopes, fears and self-awareness etc., just like a human being.

It seems to be widely agreed that a positive result *at least* entails (1). Somewhat more controversially, a considerable proportion of contemporary philosophers and cognitive scientists are inclined to suppose that (1) implies (2). This claim is based upon a behaviouristic form of functionalism—the view that mental states may be fully defined and understood, at least in principle, in terms of their objectively measurable causes and effects. For the functionalist, consciousness is as consciousness does. Of course, a facility for conversational fluency and repartee, expressed via typewritten messages, does not exhaust *all* that consciousness does—

but for the aforementioned supporters of (2), whom we might call '*Turing test* functionalists', such behaviour serves as sufficient evidence that consciousness is present.

Consciousness, however, is such an immensely rich and complex phenomenon that there is little consensus as to how it should be defined, let alone explained or understood. In order to assess the plausibility of Turing test functionalism, it may be useful to begin by turning to simpler questions—focussing, at first, upon some particular, relatively well-defined domain within the multi-faceted realm of mental life.

Let us take visual experience as an example—and consider the case of a computer system designed to process and analyse digital images and movie sequences. Suppose, for the sake of argument, that the system is just as capable as an ordinary, sighted human being in accomplishing a variety of visual tasks—discriminating colours, identifying objects, judging distances, recognising faces and so forth. Suppose, furthermore, that this system has a sophisticated natural language interface—such that within the limited domain of its visual expertise, it can answer a wide range of questions in a plausibly human-like way.

We may now devise a visual version of the Turing-test, in which referees are obliged to restrict their questions solely to the subject of visual experience. As in the standard version, human players are instructed simply to tell the truth throughout, whereas the computer is programmed to represent itself as a human being (thus, for example, it will claim, perhaps disingenuously, to have a typically human field of vision; two eyes, as opposed to a variety of interchangeable digital cameras, and so forth).

Let us suppose that the system passes this test. A succession of expert referees—including, lets say, artists, philosophers, psychologists and psychophysicists—are unable to reliably distinguish the computer from the human players. Does this tell us that our computer system has qualitative visual experience, comparable to our own?

I would like to suggest that before we decide how best to answer this question, it may be helpful to consider a slightly different case, in which we shall return to our visual variant of the Turing test—and make some adjustments.

2. Blind man's bluff

In a modified version of the visual Turing test, we shall replace the computer with a human being who has been blind from birth; claiming, unsurprisingly, never to have experienced a visual sensation. To lend support to this claim, lack of both detectable visual ability and of the brain activity normally associated with visual experience will be carefully confirmed by means of the appropriate objective tests, before the candidate is allowed to take part.

We may now proceed with our new game, which we will call, "Blind Man's Bluff". Sighted players are instructed to answer all questions truthfully, whereas blind players must try to convince the referee that they are sighted. The blind players are permitted to prepare themselves, well in advance, by studying transcripts of sessions with sighted players—thereby learning the sort of things sighted people say, when questioned about their visual experiences. They are also provided with computer terminals specially adapted for blind users, with Braille key-boards and speech-generation software to read out messages from the screen. If necessary, they are given extensive training in the use of this technology. When the game finally commences, the referee's task, of course, to distinguish congenitally blind from sighted players.

Let us suppose that at least some of the blind contestants are able to accomplish their goal. A succession of expert referees—including, lets say, artists, philosophers, psychologists and psychophysicists—are unable to reliably distinguish them from the sighted players. Does this tell us that the successful blind players have visual experiences after all?

As Turing himself might have commented, this replaces our original question: *does the Turing test provide a valid criterion for the presence of consciousness?*

3. The Lying Game

It seems clear that if, indeed, a congenitally blind player succeeds in the Blind Man's Bluff test, it will not by virtue of having subjective visual experiences, but rather, by being good at lying about them. The blind contestant can only prevail by playing what might be called a 'Lying Game'—a species of subterfuge in which a player entirely lacking a particular sort of conscious experience purports to have it, merely by imitating the phenomenological discourse of those who actually do.

We assume this must be so—because as far as we can tell, our congenitally blind players cannot have visual experiences. It might be objected that this is assumption is not absolutely *certain*. We might speculate, for example, that random nerve firings in the brain might very occasionally give congenitally blind people sudden flashes of 'inner vision'. It seems unlikely, however, this would be sufficient to provide knowledge of the rich phenomenology and structure of normal visual experience. We could test this, moreover, by initially subjecting our blind players to a Blind-Man's Bluff test *without* the opportunity prepare by studying sighted players' phenomenological reports. If they fail under these conditions, our assumptions seem fairly secure. Alternatively, we might speculate that by thoroughly familiarising themselves with the phenomenological discourse of sighted people, the congenitally blind players could somehow develop a kind of sensory empathy, to the extent that they can vividly and accurately *imagine* visual experiences. Once again, this hypothesis seems almost too far-fetched to be taken seriously—yet once again, there are straightforward ways in which we could modify our experiment in order to test it. We could monitor the brain activity of both blind and sighted players, using positron emission tomography (PET) and other techniques. If the blind players brain-scans contrast sharply with the sighted, in showing no activity in the visual cortex during the test—and if the blind players also testify, when questioned after the test, that they did not experience any visual sensations and were simply lying throughout—then our confidence in judging this to be so will be even stronger.

Now, compare the situation of the successful blind player in Blind man's Bluff with that of the successful visually-expert computer in our specialised version of the Turing test. In both cases, these players represent themselves as having conscious experiences of a particular kind. In both cases, they present the same sort of evidence: verbal responses to questions about visual phenomenology, typical of the responses normally given to such questions by sighted human beings. If we accept that the blind player almost certainly achieves this merely by mastering and mimicking the appropriate responses, then it is surely possible that the computer may have done the same. The developers may have simply programmed the computer to tell phenomenological lies—responding to questions about the subjective character of visual experience with the appropriate, typically human answers, despite the fact that there is no such subjective character to report. On the other hand, it is conceivable, perhaps, that for reasons which remain mysterious, the computer actually *does* enjoy qualitative visual experiences, just like ours. Both possibilities are consistent with our result, so the visual Turing test gives us no information, either way; it cannot tell us which of these accounts is true.

It seems clear, on reflection, that we could extend the foregoing argument to a whole variety of different types, or domains, of conscious subjective experience—and that in each case, the outcome of our *gedanken* experiment is likely to be the same. If we conclude, on these grounds, that a variety of restricted versions of the Turing test fail to serve as useful objective criteria for a the presence of particular types of subjective experience, then *a forteriori*, the standard Turing test does not provide a valid criterion the presence of human-like consciousness.

It might be objected that there is a crucial difference between our visually expert computer system and the sightless contestant in Blind Man's Bluff. The computer system has visual capabilities; our blind volunteer has none and is necessarily obliged to lie. Perhaps there is a natural law—a kind of *non-Turing*-test functionalism—which ensures that the operation of a particular type of sensory functionality confers upon its owner a corresponding type of subjective experience. However, the visual Turing test is unable to confirm whether or not this is so. We could, of course, create a new version of our computer systems with its visual capability disabled—so that its situation is equivalent to that of the blind human being. If this blind computer system were still able to pass a visual Turing test, it would be clear that this system, at least, is merely pre-programmed to lie.

In any case, the standard Turing test makes no assumptions about the sensory faculties of the computer undergoing the test. If a full set of human-like sensory faculties are required for the occurrence of sensory experience, the conventional Turing test is unable to distinguish the fully functional, sentient machine from the merely verbally adept facsimile.

It may now be helpful to set out the foregoing arguments a little more systematically. We begin with three initial assumptions:

(1) Any system that is capable of accurately emulating the typical visual phenomenological discourse of sighted human beings can pass a visual version of the Turing test.
(2) It is possible, with a high degree of confidence, to confirm that congenitally blind people never had, and cannot have, normal visual experiences.
(3) As far as know, is possible that with sufficient preparation, such congenially blind players will be able to pass the Blind Man's Bluff test.

We may take the first of these claims to be necessarily true—since normal, human visual phenomenological discourse is all that the referee has to go on, in attempting to discriminate the imposters—be they machines or congenitally blind humans—from genuinely sighted humans. We have already considered and rejected the possibility of a serious challenge to (2). The arguments that follow would be further strengthened by empirical confirmation of the possibility considered in (3), but in the absence of evidence to the contrary, the mere fact that on our present knowledge, this plausible conjecture *cannot be ruled* out is sufficient for our purposes. From (1), (2) and (3), therefore, we may now proceed as follows:

(4) As far as we can tell, the success of a congenially blind player in the Blind Man's Bluff test could only take place through mere emulation of the phenomenological discourse of sighted players.
(5) A successful result in the Blind Man's Bluff test does not reliably indicate the presence of visual experience.

We can now extend these conclusions to the general case of which the blind contestant is a particular example:

(6) A system capable of accurately emulating the visual phenomenological discourse of sighted human beings does not necessarily have visual experiences.
(7) The visual version of the Turing test is not a valid criterion for the presence of visual conscious experience.

The same considerations could also apply, of course, to other sensory domains such as hearing, taste, smell, kinaesthesia and pain; so we may generalise further:

(8) A system capable of accurately emulating typical human phenomenological discourse with respect to any particular sensory domain is not necessarily capable of undergoing the relevant subjective experiences
(9) No phenomenological-domain-specific variant of the Turing test serves as valid criterion for the presence of corresponding phenomenological aspect of experience.

In the standard Turing test, of course, the referee is at liberty to conduct all of the possible phenomenological-domain-specific tests simultaneously—together with tests of general, 'common-sense' knowledge and cognitive performance. It follows that:

(10) the standard Turing test does not provide a valid criterion for the presence of any particular sort of qualitative phenomenological experience.

We now have only to introduce one further general assumption:

(11) As far as we can tell, consciousness is always characterised by some sort of qualitative phenomenological content.

And from (10) and (11), our conclusion follows:

(12) the Turing test does not provide a valid criterion for the presence of consciousness.

4. The faith healer of deal

There are some philosophers who might be tempted to reject the foregoing conclusion by challenging (11), on the grounds that there is no such thing as 'qualitative phenomenological content'. I have suggested, elsewhere, that 'phenomenal qualities' may be straightforwardly defined as introspectible features of first person experience whose character which we find ourselves thoroughly unable to describe *formally*—i.e., purely in terms of structure and/or dynamics (Clifton 2004 [a]). Now, it seems unquestionably true that there really *are* such things; I am aware of several right now, as I write—including, for example, sensations of colour, sound, warmth, etc. I am highly confident that the reader, on reflection, will make a similar discovery; conscious experience is characterised by a wide variety of qualities which are hard-to-describe.[1] Not so, declares the radical eliminativist; those impressions that you have *of having* qualitative impressions are mere *illusions*. I like to describe this view as the 'Faith-healer of Deal position'—since its absurdity is aptly illustrated by the well-known limerick.[2] If we recall, furthermore, the two putative interpretations of a positive result in the Turing test, considered in §1, it is clear, in any case, that the radically eliminative position by no means supports the claim, which we have hitherto set out to challenge, that (1) implies (2). On the contrary, for the eliminativist, type (1) 'thinking' exhausts all that there is to consciousness and mental life; machines don't have type (2) mental states such as qualitative sensations, feelings, or beliefs—but then *neither do we*.

A more formal reply to this position than the aforementioned limerick is given in an argument I call the illusion *reductio* (Clifton 2004 [a] §4). To summarise briefly, in order to persuasively *seem like* a phenomenal quality, any hypothetically *illusory* 'impression' of a phenomenal quality must itself seem thoroughly hard-to-describe—and hence, falls under the definition of a phenomenal quality. If, on these grounds, we acknowledge the existence of qualitative phenomenological content as a characteristic feature of human-like consciousness, then our conclusion in §4 stands.

5. The functional redundancy argument

While the Blind Man's Bluff argument presented here shows that Turing test functionalism is false, it does not refute functionalism *per se*. We considered earlier, in §3, the possibility of a kind of non-Turing-test functionalism—in which it is assumed that the operation of a particular type of sensory functionality necessarily confers upon its owner a corresponding type of subjective experience. If we modified the Turing-test (and its domain-specific variants), such that in order to participate, each player must pass a series of sensory tests, then on this assumption, those players who go on to pass the test possess the relevant sort of subjective experience. We have good reason, however, to doubt this assumption, since as far as we can

[1] —Unless, of course, the reader is an artificially intelligent computer program; in which case, phenomenal qualities may not be present at all.

[2] 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.

tell, no analysis or description of a sensory functional role serves as a formal description of the associated phenomenal quality. If follows that, so far as we can judge, phenomenal qualities are *functionally redundant* with respect to the sensory and cognitive processes which they appear which to serve.[3] This is not to say that they are causally impotent *epiphenomena*,[4] but rather that the relationship between a given phenomenal quality and a particular functional role seems to be contingent, not necessary. For example, when a system such as the visually expert computer considered in §1 detects a certain patch of light of wavelength $\cong 4\times10^{-5}$ cm, we have no reason to assume that it *needs* to have an experience of phenomenal blue in order to flag the occurrence of this sensory discrimination and perform any further functions contingent upon it. On the other hand, a person who became blind during adulthood can none the less experience an impression of phenomenal blue in the *absence* of the appropriate sensory event—for example, through imagination, hallucination or dreaming.

It follows that the sensorimetric Turing test, as outlined above, is unable to tell us whether our visually expert computer-system enjoys human-like qualitative visual experiences—or none at all. The computer need not be capable of instantiating phenomenal qualities in order to perform the sensory tasks typically associated with these impressions in our own case; hence, it is able to play a lying game and pass the Turing test on false pretences. On the other hand, for all we know, the system *might*, somehow, instantiate phenomenal qualities—and indeed, it might still have access to them even if we disable its sensory functions. In that case, a visually sentient system will be falsely *excluded* by the sensorimetric Turing test.

In conclusion: the sensorimetric Turing test does not provide a valid criterion for the presence of human-like consciousness.

6. Functionalism versus consciousness

I promised at the outset to attempt a definition of consciousness. To begin with, I would like to distinguish two propositions which might be taken to guide our approach to this goal.

(1) Consciousness is a process or activity, characterised by the performance, by a system, of certain kinds of information-processing function.

(2) Consciousness is a kind of naturally occurring domain characterised by various kinds of content, of which certain properties are unique to this kind of domain.

In seeking a satisfactory definition of consciousness, which both captures and clarifies our ordinary, common-sense notion of the phenomenon this term denotes, we must consider which of these approaches serves to identify those essential features which distinguish consciousness, as such, from non-conscious phenomena—or indeed, whether both approaches should be taken into account.

The functionalist favours (1) and either dismisses (2) altogether or assumes that (2) can be subsumed into (1). On this view, consciousness consist in the performance, by a system, of one or more of some particular set of tasks such as: forming inner representations of aspects of its environment; storing such information for later recall; manipulating information creatively, to produce novel results; responding to external stimuli in organised, goal-directed ways; solving complex logical or mathematical problems; monitoring its own internal states and forming 'higher order' representations thereof; exchanging information with other systems by means of a rule-governed code or language. The *prima facie* plausibility of this approach rests largely upon the fact these are all things we can do, while conscious and (for the most part) are not so good at, while wholly unconscious. Its philosophical appeal arises from the fact that such abilities, while once considered magical or supernatural are increasingly well understood

[3] I discuss the functional redundancy argument in more detail in Clifton (2004) [b].

[4] For a critique of epiphenomenalism, see Clifton (2004) [d].

in scientific terms. The simplest of them, at least, can already be performed by machines—and we have an excellent understanding of how and why they these machines do what they do. Much progress has also been made in the development of theoretical models of how such tasks are performed in the brain—and how even the more difficult tasks could be performed by computers.

In adopting this approach, the functionalist sets out to talk about consciousness in a way which will make it easier to *naturalise*—that is, to integrate, conceptually and empirically, into our objective, scientific understanding of the physical world. A noble cause, no doubt; but it will not be served by a theory which ignores, overlooks or even openly denies the existence of the most difficult, challenging properties of consciousness experience.

An alternative approach, based on (2), can be described as *phenomenal realism*. This position is based upon the claim (defended against radical eliminativism by the illusion *reductio*) that phenomenal qualities exist—and are characteristic features of the *content* of conscious experience. Furthermore, phenomenal qualities seem unique in their (apparently) thorough resistance to formal description, in terms of structure and/or dynamics —for so far as we can tell, un-controversially *non-conscious* phenomena are always amenable to such an account—at least, to some degree of crude approximation. Indeed, all of the various cognitive functions which functionalists seek to identify with consciousness can be formally described in this way, yet no such description serves to formally describe any particular sort of phenomenal quality. Furthermore, we have plentiful evidence that at least some of these functions can occur in the absence of qualitative consciousness—and no known logical objection to the plausible view that this is possible for them all. The nature of phenomenal qualities is, at least, logically independent of the functional roles with which they are contingently associated; for no analysis or description of these roles either serves to define them or demands their existence. To *deny* their existence, however, or to ignore them (as functionalists commonly do), is to exclude from consideration the very phenomena which make consciousness unique, valuable, important, philosophically challenging—and worth having.

It will not, I hope, be considered wildly controversial, or foolishly romantic, to suggest that consciousness *matters.* According to widely held, common-sense assumptions, it is the possession of, or capacity for consciousness which makes an entity an appropriate subject for moral concern—at least, for itself and arguably, for others too. This is because the qualitative contents of conscious experience can be 'good' or 'bad', desirable or undesirable, in a sense which is inapplicable to non-conscious entities or events. That is to say the value inherent within such qualitative content is not just instrumental. Consciousness *matters*—not merely in respect of what it does, but directly, fundamentally, by virtue of the way is—the way it seems. Consciousness, in other words, is a locus of intrinsic value.

Now, I suggest, we are ready to offer a preliminary definition. Consciousness is a kind of naturally occurring domain, characterised by the presence, at least, of phenomenal qualities; some of which, at least, have the ethically relevant property of possessing, in various positive or negative degrees, subjective intrinsic value. The actual or potential possession of such a domain endows an entity with the status of being a moral subject; for what happens within that domain *matters*, at least, to the domain itself.

Given that consciousness is, in this sense, profoundly *important*, the inadequacy of the conventional Turing test as an objective criterion for its presence is a cause for considerable concern. In the case of humans, we don't have too much of a problem—for we know that the occurrence of qualitative consciousness is strongly *correlated* with the performance of certain cognitive functions and with the occurrence of certain kinds of neural activity in the brain. In higher animals, we find the same or similar cognitive functions and neural activity—so it is reasonable to assume that they are conscious too. However, we do not yet know how or why such neural activity gives rise to consciousness, in our case. This mysterious phenomenon might well depend upon particular features of the physical composition and activity of the brain which would not necessarily be duplicated in a system which appears to be functionally isomorphic. A highly detailed computer simulation of the human brain might, or might not, be conscious—and the Turing test is unable to help us, either way. Some people hope that in the future, humans will be able to upload their consciousness into such a device—but for all the Turing test can tell us, they may simply be committing suicide (Clifton 2004 [d]). Clearly, it would be desirable to have an alternative test, which will allow us to diagnose the presence or absence of consciousness within artificial systems, with a high degree of confidence.

7. The Introspection Game

I have proposed, elsewhere, an alternative to the Turing test called the *Introspection Game* (Clifton 2004 [b]). The possibility of consciousness within an artificially intelligent machine is investigated by presenting it with a variety of phenomenological questions, under what I call '*open source, open mechanism*' conditions. That is to say, the investigators are provided with complete, open-source access to the software—and unlimited freedom to examine and monitor the functioning of both software and hardware, in real time, while the computer responds to our philosophical inquiries. One of the supposed strengths of the conventional Turing test is the control against bias, for or against the possible mental attributes of a machine. We may maintain this virtue, if we consider it necessary, by conducting the phenomenological interview under the familiar conditions of the Turing test, by a remotely located referee. A second team of investigators independently examine and monitor the machine itself—with the purpose, of course, of determining whether or not the machine's replies are merely pre-programmed or imitative lies. If such a Lying Game strategy has been adopted by the system's developers, expert scrutiny of the source code should be expected to reveal it, together with some sort of 'crib-sheet'—a database of typically human phenomenological claims. Real-time monitoring of the system in operation might then be expected to show whether or not, when responding to phenomenological questions, the system merely searches its database for an appropriate reply. In that event, even if our referee identifies the machine's responses as characteristically human, we will conclude that in all probability, it merely playing a Lying Game; it does not have the kind of conscious experiences to which its phenomenological claims appear to refer.

On the other hand, if no such evidence of subterfuge is discovered and the machine nevertheless reports its awareness of an extraordinary inner, subjective world, characterised by a wide variety of formally indescribable phenomenal qualities and in particular, by states that seem fundamentally 'good', or 'bad' in themselves—then we may tentatively conclude that the machine is conscious.

The introspection game, therefore, is a valid criterion for the presence of consciousness in a machine. It may not be entirely infallible—for it is quite possible to imagine circumstances, albeit somewhat unlikely, in which false determinations, either positive or negative, might arise. Until, however, we successfully solve what David Chalmers (1995) calls the 'hard problem' and are able to precisely identify the necessary and sufficient conditions for the occurrence of consciousness in our own case, it's the best we can do.

References

Clifton, A. (2003) [a] 'An empirical case against materialism.' Unpublished MS.
Clifton, A. (2004) [b] 'The introspection game – or, does the Tin Man have a heart.' Unpublished MS.
Clifton, A. (2004) [c] 'The hazards of silicon heaven.' Unpublished MS.
Clifton, A. (2004) [d] 'Res cogitans.' Unpublished MS.
Turing, A. (1950) 'Computing machinery and intelligence.' *Mind* 59: 443-460.