The Philosophical Significance of the Turing Machine and the Turing Test

Peter Millican

Hertford College, Oxford

The concepts through which we attempt to understand the world are formed by our experience of it. Alan Turing's "imitation game" thought-experiment can be seen as an attempt to stretch that experience, and with it our concept of *intelligence*. We naturally take our own intelligence to be intimately related to our *phenomenology* – our *sentience* and *conscious awareness*. But although Turing himself sometimes evinces the same assumption, his invention of the Turing machine provides an alternative, *algorithmic* model of information processing, and thus opens the prospect – where that information processing is sufficiently sophisticated and effective to deserve the name – of achieving "intelligence" *without* consciousness.

1. Intelligence Before Turing

The objects that we find in the world appear – at least to the casual observer – to divide fairly neatly into two quite distinct categories: *purposive* and *inanimate*. We ourselves are the most immediate examples of the former, and it is only natural to take ourselves as a model for the rest. From "the inside", we both know our purposes, and self-consciously act on them. Our planned behaviour thus makes sense to us, and the actions of our family members and other humans are also explicable accordingly. Such purposive explanations are then very naturally applied further, to the animals we see behaving more or less comprehensibly in analogous ways (be they our pets, livestock, predators, birds, insects, or whatever).

Plants are less obviously purposive over a short timescale, but their growth, development and reproduction seem to manifest an equally clear teleology. Animals and plants together make up by far the majority of the most conspicuous elements of the pre-industrial landscape (something easy to overlook from within a modern house or city). It is not surprising, therefore, that before the age of modern science, the world as a whole was almost universally interpreted in terms of purpose, whether inherent or divine. Thunderstorms would vulgarly be attributed to the gods, and plagues to witchcraft. But even the academics of the time – the Medieval Schoolmen with their Aristotelian physics – took stones and stars to be as driven by purpose as animals, except that their purposes are more constant. Stones strive to reach the centre of the universe, and therefore fall to Earth; stars strive to mimic the eternal perfection of their Maker, and hence rotate around the heavens in perfect circles.

Galileo's telescope, in refuting the Aristotelian geocentric cosmology, equally sounded the death-knell for this entire picture of the physical universe. The scientific revolution that he ushered in undermined the view of inanimate objects as intrinsically purposive, replacing this with a theory of inert matter acting in accordance with rigid causal laws, being pulled and pushed around by forces and impulses that are mathematically determined by the relevant circumstances. Moving billiard balls bash into others and make them move in turn (according to their mass, angle of impact, and velocity); water gushes through a pipe under pressure from a pump; stones fall to Earth

while the moon continues to orbit, and we discover that both the falling and the orbiting can be neatly explained by Newton's postulation of a force of gravity that attracts according to an inverse square law. Modern physics significantly complicates this picture, of course, interlinking space with time and introducing an element of indeterminism. But the general mechanistic paradigm remains, the future of inanimate things unfolding from their past through mathematical laws that are purposeless and oblivious of any final destination.

From this modern perspective, the idea of purpose in inanimate things seems puerile and superstitious, or an occult relic of a pre-scientific era. Even most examples of living organisms, including plants and "lower" animals, cease to be purposive, their appearance of teleology *explained away* by Darwinian selection. Genuine purpose lies exclusively in the domain of *conscious* beings, *desiring* certain ends and – at least in the case of humans and privileged "higher" animals – *thinking* about the means to achieve those ends before acting accordingly. To think in this way to good effect is to be *intelligent*, a concept which thus ties together *conscious purpose* with the *effective processing of information* to identify the means to a desired outcome. Without the desired outcome, there would be no target for the information processing. But in the assessment of intelligence, it is the effectiveness of that processing rather than the strength or nature of the desire that provides the crucial measure. Human beings are the pre-eminent intellects of the natural world not because our desires are stronger than those of, say, a dog, but because we are so much better at identifying unobvious patterns, forming sophisticated plans, and calculating complex consequences.

2. Turing Machines, Intuition Pumps, and a Word of Caution

In his famous paper "On Computable Numbers" (1936),[1] Alan Turing came up with a precise model of an information processing machine – now universally known as a Turing machine – and provided an informal argument to suggest that when suitably programmed, this could faithfully execute any well-specified algorithmic process that can be carried out systematically by a human thinker. Suddenly a new question arises: *Should such information processing, as performed by an inanimate machine, be deemed genuinely intelligent?* Our experience of nature has not prepared us for this question, for although we have learned to think of *intelligence* as primarily a measure of the sophistication of information processing, we have also understood it as confined to *conscious* beings, planning how to achieve their ends. Now we are in a novel situation, faced with a machine which is clearly capable of processing information – of *calculating* answers to the sorts of questions that we standardly think of as demanding intelligence – and yet which has no ends of its own, and whose functioning has no need of *reason* as traditionally understood: no need of genuine *understanding*, *insight*, or *consciousness*.

Philosophers attempting to circumscribe the boundaries of some controversial or troublesome concept often appeal to thought-experiments, nicely characterised by Daniel Dennett (1995) as "intuition pumps". Given the context described above, two sorts of thought-experiment naturally suggest themselves. On the one hand, the advocate of machine intelligence can point to some suitably impressive example(s) of the sophisticated information processing achievable by an appropriately programmed Turing machine, and ask: *How can something which achieves <u>this</u> be denied genuine intelligence?* On the other hand, the opponent of machine intelligence can emphasise

[1] The paper is technically daunting, but is presented and explained very effectively by Petzold (2008).

the crude, mechanical basis of the entity which is performing the processing, and the trivial individual steps by which it is operating, and ask: *How can anything which works like this be judged genuinely intelligent?* Turing himself takes the first path, presenting us in his paper "Computing Machinery and Intelligence" (1950) with a scenario – a successful "Turing test" – in which it seems unreasonably chauvinistic to deny intelligence.[2] John Searle, with his well-known "Chinese room" thought-experiment, takes the second path, focusing not on the outcome but on the method of processing, which seems so fantastically divorced from reality as to lack any semantic grounding.[3] The operator in Searle's room cannot plausibly be considered as reasoning *about* whatever is represented by the Chinese symbols he manipulates; hence the processing he carries out cannot be classed as *genuinely* intelligent. That, at least, is the moral that Searle would apparently have us draw.[4]

Thought-experiments designed to elicit particular "intuitions" can be endlessly seductive for philosophers, but a severe note of caution is appropriate. Consider, for example, the following argument:

> "Performance at chess cannot provide an adequate criterion of intelligence, even of a domain-specific kind. For suppose that someone were to write a computer program of only a few dozen lines of code (in a standard general programming language), which could play chess at a grandmaster level in real time. Such a crude program could not possibly count as genuinely intelligent. Hence grandmaster performance at chess is not a reliable proof even of intelligent chess-playing."

It might well be true that we would be reluctant to count such a short computer program as "genuinely intelligent". But of course the fundamental hypothesis of this thought-experiment – that such a program could possibly play grandmaster chess in real time – is utterly ludicrous. So we have no reason for taking it seriously as a guide to the boundaries of our concepts. Indeed it is easy to see that were we to allow this sort of thought-experiment quite generally, it could without further ado rule out *any* performance-based criterion of intelligence.[5] But this seems outrageously simplistic,

---

[2] Turing calls this the "imitation game", but it is now universally known as the "Turing test" (at least when it involves a human interrogator interacting by teletype machine with one other human and one computer). In general terms, a computer program "passes" the Turing test if it maintains a text-only conversation with sufficient human realism that the human interrogator cannot reliably distinguish between it and a human conversationalist. More detailed aspects of the test are discussed in §6 below.

[3] The most familiar Chinese room scenario (Searle 1984, p. 32) involves a conversation conducted in written Chinese by means of cards posted into and out of a room, where the incoming cards express meaningful questions, and the outgoing cards provide meaningful and appropriate answers to those questions (such as might be produced by a competent and intelligent native speaker of Chinese). The twist is that the man inside the room has no knowledge whatever of the Chinese language or of the *semantics* – the meaning – of the symbols he is reading or writing. Instead, he is generating his written "answers" by strictly applying rules based purely on the *syntax* – the shape and structure – of the "question" character strings that he receives, these rules being specified in books contained within the room. Searle wishes us to conclude that the apparent meaningfulness of the answers that the man generates is an illusion, a conclusion which can then be taken as equally applicable to conversations generated by natural language processing computer programs.

[4] I say "apparently", because although Searle presents his argument as an attack on "strong artificial intelligence" and on the idea that machines can "think" (e.g. Searle 1980, p. 417; 1984, p. 36; 2002, p. 56), he generally expresses his thesis not as a denial of *intelligence* but rather of "intentionality", "cognitive states" (e.g. 1980, p. 417); "a mind", "mental states" (e.g. 1984, p. 37); "cognitive processes", "mental content", "semantic content", or "consciousness" (e.g. Searle 2002, §I). Since my focus is on Turing I shall not address this issue in detail here, but see note 12 and §§4-5 below.

[5] "Suppose that someone were to write a computer program of only a few dozen lines of code … which could solve any problem of kind *X* …". And so forth.

given that our notion of intelligence is, as mentioned above, one that we tend to assess primarily in terms of information-processing performance.

Note that performance here involves issues of resources, as well as output. It is not terribly difficult to write a computer program of modest length that plays infallible chess, if time and memory space are no object: simply analyse every possible line to the end, and score each as checkmate or as drawn (by stalemate, repetition, or the 50-move rule), chaining back accordingly. But such a program, whose first calculated move will require vastly more steps than there are picoseconds in the entire history of the universe, is unlikely to be deemed "intelligent". Much the same applies to Ned Block's "Blockhead" program, supposed to be programmed in advance with pre-prepared "intelligent" responses for every conceivable sequence of verbal inputs in a conversation of some pre-defined length.[6] Such a program is possible only "in principle", for even to set it up to cope with a fairly short conversation could consume more memory than the universe can hold. It seems entirely reasonable to deny that such programs can count as genuinely *intelligent*, when their mode of operation is so far removed from the clever techniques that intelligent organisms have evolved to enable us to negotiate our way through complex problems with very limited resources.

Searle's "Chinese room" combines outrageous unfeasibility with elements of genuine impossibility, because it hypothesises that intelligent answers – as good as those of a typical native Chinese speaker – could be framed by following purely syntactic rules in a context where the operator of those rules has no means of taking account of a changing world, both external and internal, whose events form the subject-matter of so much of our conversation. When asked (the Chinese translation of) questions such as "Do you like the weather we've been having?", "Did yesterday's news about *X* upset you?", "How many times did I knock on your door just now?", or "Have you disagreed with anything I've said in the last five minutes?", the operator's syntactic rules give no scope for sensory input, real-time updating, or emotional reaction, and so however sophisticated those rules might be, he cannot possibly match the response of someone who understands the question. But even if the questions are artificially limited to comprehension of a fixed story written in Chinese, rather than being interactive,[7] the suggestion that intelligent responses to arbitrary Chinese questions could be generated by Searle's specified method – through the manual consultation of purely syntactic rules recorded in books within a room – is as ridiculous as the idea that a 50-line computer program might play grandmaster chess in real time. "Surely", Searle's scenario implicitly urges, "something that operates in such a manner cannot possibly be deemed intelligent". This indeed seems persuasive, but then something that operates in such a simplistic manner could not possibly reach the level of performance that Searle is postulating, so the significance of his thought-experiment is crucially undermined.[8]

---

[6] See Block (1981). It is unclear who first coined the nice name "Blockhead" for the program described in this paper.

[7] As in the original 1980 version of Searle's Chinese room scenario. For some useful background to that article, see Preston (2002), pp. 16-19.

[8] Searle might respond that it is *conceivable* that something operating by the Chinese room method – even under realistic constraints of space and speed – might achieve the required level of real-time performance, and this should therefore be considered *possible*. However it is obvious from cases such as the provability (or otherwise) of Goldbach's Conjecture that conceivability can be taken as a reliable guide to possibility only, at best, where it is interpreted as involving *clarity and distinctness* of a fairly strong kind. I can of course conceive in a general sense (e.g. sufficient for understanding the words) what it would be

3. Turing Machines, New Paradigms, and Open Texture

I started by suggesting that our modern concept of *intelligence* was established within a world apparently divided between two main categories of entities. We ourselves, together with "higher" members of the animal kingdom, are organisms moved by conscious desires, able to process and exploit information (with various degrees of sophistication) in the conscious attempt to fulfil those desires. All other things lack consciousness, and therefore cannot be moved by such desire, nor apprehend information. These unconscious entities include physical objects, which act mechanically: pushed or pulled around by impacts and forces that are blind to any final outcome. Plants and "lower" animals, though presumably equally unaware, behave in ways that seem to manifest purpose, sufficiently so that for millennia it proved almost irresistible to attribute this behaviour to the influence of a divine being with human-like intentions. Darwin's theory of evolution was revolutionary not only because it removed the need for such a designer-god, but more fundamentally, because it introduced an entirely new mode of explanation which was neither mechanical nor purposive. Such explanatory innovations are rare but momentous: other examples would be the development of mechanism itself by Galileo and others (e.g. Descartes and Boyle) in the seventeenth century, and its challenge by quantum mechanics in the twentieth century, which brought a very different conception of physical explanation.

The invention in 1936 of what we now know as the Turing machine bears comparison with these paradigm shifts.[9] For it provides a way of specifying processes in *algorithmic* terms that are neither purposive nor mechanical, but have common features with each. Like a purposive explanation, an algorithm is couched in terms of the abstract processing of information. But like a mechanical explanation, this processing is at bottom "mindless" and automated, taking no account of any semantic significance and paying no regard to any desired endpoint. This – I suggest – is what opened the possibility that Turing presented so forcefully in his 1950 paper, of intelligent information processing that is automated rather than purposive. Before the Turing machine, information had to be understood in terms of its significance to a *conscious mind*. But Turing saw that information – and information processing – could be understood quite differently, thus opening the possibility of machine "intelligence" gauged in terms of inputs and outputs rather than requiring any sort of internal understanding. Hence we reach the idea of a thought-experiment that compares the external behaviour of man and machine, judging the latter to be intelligent if it can do equally well. As we saw in the previous section, such thought-experiments need to take account not only of inputs and outputs, but also the constraints of our practical situation. Even the best-equipped organisms are limited in knowledge, capacity, time, and other resources. This puts a premium on the effective exploitation of our limited means, on

---

for a 10-line computer program to play infallible chess in real time. But I cannot clearly and distinctly conceive how such a program would operate, and it is *very obviously* not a genuine possibility. For more on this, and on Hume's influential appeals to the Conceivability Principle, see Millican (forthcoming) §5.

[9] Floridi argues that "the best way to understand the information turn is in terms of a fourth revolution in the long process of reassessing humanity's fundamental nature and role in the universe. We are not immobile, at the centre of the universe (Copernicus); we are not unnaturally distinct and different from the rest of the animal world (Darwin); and we are far from being entirely transparent to ourselves (Freud). We are now slowly accepting the idea that we might be informational organisms among many agents (Turing) …" (2008, p. 651). Whether or not one accepts this account (e.g. I would be inclined to replace Freud with Hume and the development of cognitive science, cf. §4 below), it is interesting that the paradigm shifts I have identified – in terms of the discovery of new modes of explanation – correspond quite closely to major upheavals in our understanding of our place in the universe.

efficient and flexible processing with uncertain inputs and under pressure of time. We naturally judge intelligence accordingly, and deprecate the inefficient brute force methods of the implausible thought-experiments of Searle and Block, which lack any practical utility and differ so radically from any familiar reality.

Turing's "imitation game" thought-experiment, however, still remains to be judged, and we surely know now – even if this was hard for most of his readers to appreciate in 1950 – that the level of computational linguistic ability that it postulates is relatively plausible. Over the last decade, interactive computer systems have made huge strides in the processing of natural language (as illustrated, for example, by the development of automated translation systems), and although they still have a long way to go, it is by no means obviously ridiculous to consider a future system – even within the next few decades – that might achieve something like the level of performance anticipated by Turing. Admittedly some aspects of his thought-experiment are less plausible than others, notably his requirement that the envisaged system should be able to pass for a human in general conversation, informed as this might be by personal emotions and by reference to changing events.[10] So to ensure that our discussion remains solidly grounded in foreseeably plausible reality, let us suppose only that the challenge of the Turing test *has* been fulfilled in an extensive – though not unlimited – factual domain: perhaps the science of chemistry. Suppose that computers have been programmed in such a way as to be able to sustain long and detailed conversations, appropriately directed, with complex, accurate reasoning and interlocking themes, apparently well-informed about all relevant aspects of chemistry.[11] Should we call such conversational behaviour "intelligent"?

In making this judgement, there is no reason why we should confine ourselves within narrow behaviourist limits, and it is entirely legitimate to take account of obvious points regarding the nature of any such program. Clearly such sophisticated discourse about chemical interactions will have to be informed by representations of molecular structures and relevant laws and forces: this is not a Blockhead-style lookup table, nor a "chatterbot" designed to mislead (cf. §6 below). Real information processing is taking place, generating appropriate and informative responses by reference to the same mathematical and structural relationships that would inform a human expert,[12] but none

---

[10] Note, however, that Turing would "wish to permit every kind of engineering technique to be used in our machines" (1950, p. 435), including "the best sense organs that money can buy" (p. 460). Consideration of sensory input plays a large role in his discussion of learning machines (pp. 454-60), and presumably explains why he expresses a special concern about the possibility of extra-sensory perception, which (were it to occur) could not be replicated mechanistically (pp. 453-4). Searle standardly restricts his operator to information gleaned from books inside the Chinese room, but he takes the message of his thought experiment to apply equally to a robot equipped with appropriate sensors (1984, pp. 34-5)

[11] This enables us to put aside the question of whether such a program could convincingly discuss matters that arguably require essential reference to human perceptions or emotions, such as sensory phenomena, morals, or aesthetic appreciation. The presupposition that intelligence in *one* area does not require intelligence in *all* seems highly plausible to me, but could perhaps be threatened if, for example, it turned out that only a "global workspace" could solve the frame problem (cf. Shanahan and Baars, 2005).

[12] Searle might contest this, on the ground that there is no *semantic* connection between the representations in the program and the real-world features they represent. But for present purposes, the fact that there is a well-designed isomorphism between the relationships as understood by the scientist, and those formally manipulated by the program, will do. Clearly in some sense there is *information* being processed, even if that information fails to live up to Searle's "semantic" requirements. Space does not permit further discussion of Searle's concerns here, but suffice it to say that I consider his notion of the "semantic" to be fundamentally obscure, and liable to dissolve under close analysis. As the references in note 4 above indicate, his terminological promiscuity tends to conflate *information processing* with

of it – at least on the part of the program – is the least bit *conscious*. Does this then debar it from deserving the accolade of *intelligence*?

I would like to suggest that we cannot necessarily expect an unambiguous answer to this question, because it concerns the application of a concept beyond the context for which it has evolved. Our common-sense world-view seems to imply a general division between things that are consciously purposive and calculating, and others that are neither conscious, nor purposive, nor calculating. So it is not surprising that we then find it hard to classify a novel kind of entity which seems to calculate very effectively (in a sophisticated manner, and to a useful purpose), but which itself entirely lacks any kind of consciousness, and hence lacks any awareness or "internal" understanding of either the apparent purpose or the calculation.

We have here a case of what Friedrich Waismann called *open texture*. A concept or term is said to be *open textured* if our understanding of it does not "provide in advance for all possible cases".[13] Our concepts are framed, or adapt, to fit the circumstances in which they are standardly employed, and they commonly fail to have determinate criteria of application in abnormal, unanticipated circumstances. Suppose, for example, that marriage is defined as being allowed only between a man and a woman, in a society in which it is absolutely taken for granted that everyone has an unambiguous sex (and gender) throughout their life. This rule might seem to be entirely clear and precise; indeed those who frame it take it be so. But it can nevertheless become indeterminate if, for example, someone is born chromosomally male but physically female, or if sex-change operations occur. As the philosopher of law Herbert Hart insisted and this example illustrates, open texture is particularly important in legal contexts, which often hinge on the precise boundaries of rules that have been specified without even considering, let alone defining, their application to "all possible cases".[14]

4. Intelligence and Consciousness

If the concept of intelligence is open-textured in this way, then its application to suitably programmed computers is an open question, rather than one we should expect to be able to decide by simply analysing our existing concept. But this does not imply that the question has no best answer, and we have already seen at least one good reason for siding with Turing rather than Searle. For although we standardly take an intelligent entity to be one that has *conscious awareness and purpose* as well as *effectiveness in processing the relevant information*, nevertheless when we judge one person to be *more intelligent* than another, we do so almost exclusively in terms of the latter criterion. Thus we do not typically consider mathematical brilliance to be any sort of measure of

---

phenomenology, whereas I shall argue in §§4-5 below that these are best distinguished, since "intelligent" processing does not necessarily require *consciousness*. Once distinguished, the plausibility of Searle's claim that a computer program could not possibly have "semantic" relationships – at least in the *information processing* sense – is significantly weakened (especially if we consider the possibility of directly connecting the program to reality through appropriate sensors and manipulative mechanisms). Moreover even if the claim were to be accepted (e.g. on the basis that any fully adequate semantic relationship must involve *conscious* intentionality), it would then require a further argument to move from a lack of *semantics* to a lack of *intelligence*.

[13] Quoted from Williamson (1994), p. 90. Williamson discusses Waismann's use of open texture on pp. 89-95.

[14] An entertaining example is provided by a famous *Punch* cartoon (6 March 1869, p. 96), in which a railway porter is telling an old lady about the price of travelling with her menagerie of pets, given rules which specify a cost for dogs only: "Station Master says, Mum, as Cats is 'Dogs', and Rabbits is 'Dogs', and so's Parrots; but this 'ere 'Tortis' is a Insect, so there ain't no charge for it!"

the quality of a mathematician's inner life – the motivational desires, feelings of effort, or even poetic urges that he might experience whilst proving his next theorem. All this subjectivity is irrelevant, and it is the objective quality of his proof production that dominates, *except* in so far as we are inclined to require *some* inner life before we are prepared to count "him" as a mathematician at all (as opposed to a mathematical tool).

This consideration can be pushed further, by noticing that for humans, at any rate, there is often an inverse relationship between these subjective and objective qualities. Perhaps the most familiar example is in driving a car, where the seasoned expert achieves high performance with little focused "consciousness" of what he is doing – or subsequent memory of having done it – while the stumbling learner driver is only too conscious of every tense observation and manoeuvre. In the same way, the novice chess-player struggles to find a good move, vividly aware of his efforts and uncertainties, reflecting carefully and anxiously on all the considerations that come to mind. The grandmaster, by contrast, typically finds his move effortlessly, almost without conscious thought and entirely without struggle; moreover when asked to explain his "thinking", he might well have nothing better to say than that "in this sort of position, *that* is obviously the right move to play". Here again our common-sense identification of *intelligent* information processing with *self-conscious* information processing is contradicted, as we find that greater expertise is frequently accompanied by *less*, rather than more, articulacy.[15] And accordingly the person who has had to struggle to acquire a skill often makes the better teacher, having reflected far more on what the skill requires, and able to relate more closely to the difficulties of students. But we do not on this account judge him to be the better practitioner of the skill, even where the skill is one that we think of as paradigmatically "intellectual". Executing an intellectual task is one thing; reflecting on it quite another.

Pushing even further in the same direction, it turns out that there is little correlation between the sophistication of information processing that common tasks require, and their typical psychological impact or effort. Indeed it seems that the vast majority of the most complex processing that takes place in our brains is entirely unconscious, and remarkably little of our mental life can properly be explained in terms of reflective reasoning and explicit inference. David Hume famously pioneered this message, proving the impossibility of accounting for such basic mental operations as inductive inference or the identification of persisting objects in terms of any traditional concept of reason. Though we might suppose that we are transparently *apprehending* rational connexions between past and future, or passively *perceiving* continuing things through time, in truth our minds (or at least our brains) are actively supplying crucial contributions of their own. It is these active inputs that enable us to move to conclusions beyond what pure reason would warrant, and to smooth over irregularities in the flux of sensations. And because they are creative rather than cognitive processes – reading *into* the world of our experience rather than *off* it – Hume attributes them to "the imagination" rather than to "reason". The same lesson has been emphasised even more in recent years, with discoveries prompted by studies in artificial intelligence. It is now clear, for instance, that even the identification of objects in a visual scene *at a single time* essentially involves active processes of edge detection, shadow interpretation, and so forth, all of which are typically *subcognitive* and therefore unavailable to consciousness. And this increased appreciation of the sheer computational complexity of everyday cognition has gone together with a re-evaluation

[15] See Michie (1993) and also my introduction to Millican and Clark (1996), pp. 2-3.

of the familiar examples of "intelligence" that once seemed to represent the pinnacle of intellectual achievement. Arithmetic, for instance, seems abstract and difficult for humans, and is hard to master without years of schooling and practice. Yet compared to the computational difficulty of, say, tracking and catching a ball whilst running (something which many of us can do with relative ease, and which dogs seem to do quite naturally), arithmetic is utterly trivial. Again the lesson seems to be that if we wish to preserve the criterial correlation between *intelligence* and *competence founded on sophisticated information processing*, then we must be prepared to cast off the folk-psychological assumption that greater intelligence requires greater consciousness of what we are doing. With that assumption discarded, there is much to be said for relinquishing the requirement of consciousness entirely.

5. Information Processing and Phenomenology

Throughout this discussion I have resisted any *conflation* between "intelligence" and "consciousness", whilst fully acknowledging that our naïve concept of the one significantly implicates the other. This is important, because discussions of the Turing test are often horribly muddied by a failure to distinguish two quite different features of what we take to be intelligent thought, namely the *information processing* that it involves, and the *phenomenology* that potentially accompanies that information processing. Too often, the possession of *intelligence* is conflated with possession of a *mind*, yet it seems to me that the connotations of the two are radically different. When we consider an entity as having a mind, the crucial factor is not so much the quality of its intellectual processing as its possession of an "inner life", or as Thomas Nagel (1974) famously put it, there being "something it is like" to be them. When we say that we are "minded" to do something, we are expressing a *felt desire* rather than any intellectual process. And nothing said above has given the slightest ground for supposing that an electronic computer – no matter how cleverly it might be programmed – is able to experience genuine feelings. I have argued that we should be prepared to accept the notion of unconscious *intelligence*, but there is no such compelling reason to countenance unconscious *desires*, let alone unconscious *feelings*.[16] Some might wish to do so, attracted either by exotic Freudianism or, at the other extreme, by the austere objectivity of behaviourism or functionalism. But there is no need for us to adjudicate on these things here, and I am happy to allow the opponent of machine intelligence to insist that even the merest *wish* – let alone a *passion* or a *craving* – is something that essentially requires *feeling*, and hence is confined to conscious beings.[17] We have already noted that this does not prevent *unconscious* beings from exhibiting *apparent* teleology, as we find in much of the animal kingdom and universally amongst plants. But again, for present purposes, I am quite happy to allow Turing's opponent to explain this away by the familiar appeal to Darwinism, and to reserve the term *desire* for the genuine (i.e. conscious) article. Such a reservation, however, is entirely consistent with allowing the possibility – indeed the manifest reality – of unconscious *intelligence*.

---

[16] Unconscious sensations may provide an intermediate case, in that although *sensory awareness* can be seen as a source of information (and to that extent abstracted from phenomenology), conceptually it seems to be tied more closely to its internal Nagelian character than in the case of *intelligence*.

[17] So here I am content to agree with Searle in opposing the view that "any physical system whatever that had the right program with the right inputs and outputs would have a mind in exactly the same sense that you and I have minds. … that it must have thoughts and feelings, because that is all there is to having thoughts and feelings: implementing the right program." (1984, pp. 28-9).

Turing himself, unfortunately, is guilty of the conflation that I am resisting, and perhaps deliberately so. In his 1950 paper he considers Geoffrey Jefferson's "Argument from Consciousness" as a "denial of the validity of our test", and his response is to compare it with solipsism, as though we could have no better reason for denying consciousness to a (suitably conversing) computer than we have for denying it to our fellow humans. But this response is weak, and should convince nobody who takes consciousness to have an ontological reality over and above behaviour and functional role. Certainly the subjectivity of consciousness seems mysterious, and perhaps all the more so as our psychological and physiological science has become more objective. The relationship of consciousness to our physical brain is hard to make sense of, as is its evolutionary function: even if we ignore the difficulty of understanding how consciousness can arise from physical matter, it remains obscure how, having arisen, it can contribute to our biological success (as the popularity of thought-experiments involving "zombies" testifies). Nevertheless, if there is one solid certainty in all this,[18] it is that consciousness must indeed bring some such evolutionary benefit, perhaps by facilitating a more efficient form of perspectivally-informed processing than would otherwise be possible (e.g. spatio-temporal, perhaps, or in terms of our ability to employ a theory of mind about our fellows).[19] And that being so, we have every reason to suppose that the same biological make-up which generates our own capacity for consciousness does exactly the same for others of our species (and, indeed, of similar species).[20] No such argument can be made in the case of a programmed computer or robot. On the contrary, such a machine's behaviour – however closely it may be designed to mimic our own – is precisely explicable in terms of its program: that is what the program has been designed to do! When the machine produces an output which, in a human, would be expressive of consciousness, we know that the reason it does so is that it has been programmed appropriately (even if the detailed algorithmic mechanism is unpredictable or too complex for us to discern). Genuine, full-blooded, ontological *consciousness* – whatever exactly that might be supposed to be beyond behaviour and functional role – is an entirely gratuitous postulation in such a case, eliminable immediately with a slash of Ockham's Razor. So Turing's anti-solipsistic move is powerless against someone who insists that we have such consciousness.

Note that this argument does not depend on the assumption that genuine consciousness is irrelevant to the achievement of sophisticated information processing; nor would it be refuted by the discovery that some forms of information processing are entirely beyond the practical capacity of anything that lacks a conscious perspective. For the latter discovery could only plausibly be made in respect of a form of information processing that had *not* been achieved by a computer. If a computer were to achieve it, that would *ipso facto* provide overwhelming evidence that the computer's

---

[18] Such certainty is vastly more likely to be found in reasoning based on scientific considerations that are liable to empirical test – or on formal rules whose reliability can be rigorously tested by mechanical application to numerous cases – than in the aprioristic (and typically ad hoc) untestable argumentation of armchair philosophers. Anyone disinclined to accept this Humean point (1748, 12.27-9) would be well advised to ponder the track-record and shifting fashionable tides of philosophical armchair speculation!

[19] If consciousness had no causal impact on behaviour, but just somehow arose as an epiphenomenon, then it would be a complete coincidence that such a manifest correlation has evolved between subjective feelings and bodily events. If the subjective pain of banging my knee, or the pleasure of tasting honey, are causally inert, then there is nothing to tie them evolutionarily to the events that characteristically generate them, and from the point of view of survival, they could just as well be reversed.

[20] For an illuminating discussion on the connection between evolutionary considerations and the inference to mental states of others, see Sober (2000).

programmed powers were sufficient for it, and thus count decisively against the hypothesis that anything more was required.

6. Evaluating the Turing Test: The Lessons of ELIZA

In considering the significance of Turing's thought-experiment, we should bear in mind the state of computer technology – both hardware and software – at the time when he came up with it. He could not point, as we can now, to sophisticated computer systems achieving feats of information processing hugely beyond the powers of the unaided human brain, not only in relatively abstract calculation (such as arithmetic or chess-playing), but also across a large and ever-increasing range of scientific enquiry. What he sought, therefore, was not a general criterion of intelligent behaviour, but a clear illustration of one sort of behaviour that anyone *would* recognise as paradigmatically intelligent were it to be achieved.[21] And in order to make this illustration relatively plausible within a reasonable timescale, that behaviour had to be confined to *verbal* interaction. In this context, his choice of test was judicious, his examples convincing, and his predictions remarkably accurate. Here, first, is an example of a conversation from the 1950 paper which, if spontaneously produced (and hence not pre-arranged in any way), would surely tend to persuade us that the Witness is capable of responding to such questions appropriately and "intelligently":

> "Interrogator: In the first line of your sonnet which reads 'Shall I compare thee to a summer's day', would not 'a spring day' do as well or better?
>
> Witness: It wouldn't scan.
>
> Interrogator: How about 'a winter's day' That would scan all right.
>
> Witness: Yes, but nobody wants to be compared to a winter's day.
>
> Interrogator: Would you say Mr. Pickwick reminded you of Christmas?
>
> Witness: In a way.
>
> Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.
>
> Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas." (1950, p. 446)

However Turing is not so rash as to predict that performance at anything like this level is likely to be achievable soon. At the beginning of §6 of his paper, "Contrary Views on the Main Question", he famously makes the following far more modest prediction:

> "I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well than an average interrogator will not have more than 70 per cent. chance of making the right identification after five minutes of questioning. … I believe that at the end of the century the use of words and general educated opinion will have altered so much than one will be able to speak of machines thinking without expecting to be contradicted." (1950, p. 442)

The standard he suggests – whereby an *average* interrogator is supposed to be able to do no better than distinguishing between the computer and a human with *70% accuracy* after a mere *five minutes* of questioning – is not particularly high.[22] Had this been a

---

[21] Turing makes clear that he is seeking a *sufficient* test of intelligence rather than a *necessary* condition, when addressing the objection: "May not machines carry out something which ought to be described as thinking but which is very different from what a man does?" (1950, p. 435).

[22] This would mean that a sequence of 100 members of the public, each faced with the task of distinguishing a human from the computer through a five-minute "interrogation" of each of them, could

serious goal of artificial intelligence research, I expect that it would have been solidly achieved by the year 2000. As for the other part of Turing's prediction, by the end of the century it had indeed become fairly commonplace to talk of computers "thinking", especially about difficult information-processing tasks taking place in real time. Other psychological verbs have also become natural to apply to computer programs, with minimal if any embarrassment, and a conversation like this about a chess-playing program would not seem out of place:

> "Why is the computer taking so long to respond to your queen move?"
>
> "It's thinking hard, because it's realized that if it tries to defend against my attack by bringing its knight over to protect the king, I'll be able to grab its pawn on the other side. It's displaying now that it assesses the position as better for me materially, but it seems to be predicting that it can get some activity to compensate if it decides to let the pawn fall."

No doubt many philosophical pedants, hearing this conversation, would want to insist that the psychological verbs are being applied only loosely or analogically. But the fact remains that such application is extremely natural, and by now rather likely to be used "without expecting to be contradicted". We have here symptoms of precisely the sort of conceptual evolution advocated above, whereby increased habituation to a changed reality leads to a corresponding adaptation of our traditional concepts.

This conceptual evolution, however, has not been significantly fostered by work towards satisfying the Turing test, which has led in a very different and less productive direction, namely the development of "chatterbots" that are typically at best amusing curiosities rather than serious tools. Indeed with hindsight, it is a shame that Turing not only proposed his "imitation game" as an illustration of how a computer could manifest intelligence (as in the sonnet conversation above), but also gave the impression that it could provide a rough measure of success in developing machine intelligence. For his quantitative prediction – that a 30% success rate at impersonation over five minutes' questioning would be achievable by the end of the century – naturally suggests that a higher success rate, over a longer period of questioning, would be a suitable indicator of progress. But unfortunately, it is no such thing, because as the experience of Joseph Weizenbaum's ELIZA program of 1966 quickly showed, success in impersonation can be far more readily achieved by trickery and misdirection than by genuinely intelligent processing. Hence we have the sad spectacle of the Loebner Prize organisation, year after year, holding a Turing test challenge which encourages the further development of these largely pointless chatterbots, whose programmers, rather than attempting to write systems that handle language intelligently, instead focus on marginally improving their methods of *appearing* to contribute appropriately to the flow of a conversation. What these programs actually do is typically nothing more intelligent than recognising simple patterns in the input and replaying adaptations of the interrogator's own phrases, suitably modified by switching first- and second-person etc. Where even this fails, they say something entirely neutral (e.g. "GO ON"), revert to some previous input, change the subject, or disguise the failure by saying something that makes no pretence of continuing a thread of conversation but is instead designed to appear amusingly eccentric (a manoeuvre which can even be presented positively as giving the chatterbot

---

be expected to achieve 70 correct identifications, and 30 incorrect. So there is no requirement here that the interrogators are deceived most of the time. It is perhaps better to think of this performance as 30% *successful impersonation* by the computer program, rather than 70% *accuracy*, since the point of the test is to provide a measure of the *program's* achievement, rather than the *interrogators'*. (But note that 50% successful impersonation would be enough to imply indistinguishability from a human – any more than that, and the program is implicitly being judged to be "more human than a human".)

a distinctive "personality"). The original conversation in Weizenbaum's 1966 paper, which introduced most of these techniques, includes such gems as:

> "Men are all alike. [pattern: 'alike']
>
> IN WHAT WAY
>
> …
>
> Well, my boyfriend made me come here. [pattern: 'my X' – this also prompts saving of X]
>
> YOUR BOYFRIEND MADE YOU COME HERE
>
> …
>
> I need some help, that much seems certain. [pattern: 'I need X']
>
> WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
>
> …
>
> My mother takes care of me. [pattern: 'my mother X']
>
> WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
>
> …
>
> Bullies. [no recognized pattern, so previously saved 'my X' phrase is re-used instead]
>
> DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE"

Such a conversation can indeed *seem* intelligent, because observers are typically unaware of how much its apparent coherence depends on trickery and pretence (or, in the last case, sheer coincidence), and so they systematically give the computer the benefit of the doubt by interpreting the outputs as maximally appropriate within the conversation. Weizenbaum also cunningly has his program ELIZA play the role of a Rogerian psychotherapist, whose method consists largely of echoing back the human user's own thoughts, in order to elicit further such thoughts.[23] This ruse is quickly exposed if one sets two clones of such a program conversing with each other: without any injection of substantial content from human input, the conversation soon descends into aimless vacuity.

Since Turing presented his test as a demonstration of how computers could in principle manifest human-like intelligence, it is ironic that the main objection to the test is how *unintelligent* humans can be, both as conversationalists and as interpreters. In some areas of ordinary life, a fair amount of human conversation can consist of vacuous responses which engage only vaguely with what has gone before. Presumably for this reason, our interpretation of others' contributions can be excessively uncritical and over-generous, even when we have knowingly been put in the role of an "interrogator" judging their intelligence. So although human conversational behaviour is generally intelligent up to a point – and often highly so – it is hardly a *paradigm* of intelligence, and there is no reason why *indistinguishability from a human* should be seen as the ideal criterion of intelligence, let alone *indistinguishability as judged by an average human* (which is thus doubly polluted by human sloppiness and fallibility).

None of this is to deny Turing's claim that genuine indistinguishability over an extended and suitably probing discussion by a discerning interrogator – as illustrated by

---

[23] For an implementation of Weizenbaum's "DOCTOR" script (based closely on the appendix of his 1966 paper and generating the dialogue he quotes), within a fully documented learning environment that facilitates practical chatterbot development and experimentation without requiring programming expertise, see www.philocomp.net/ai/elizabeth/.

his sonnet conversation – would provide a reasonable basis for ascribing intelligence (at least on a provisional basis).[24] But were anything remotely like this to be achievable in practice, the main problem with the test would become not the over-generosity of uncritical interrogators, but rather their excessive discernment. For in this situation, the main concern of programmers attempting to fulfil the test would be not the maximising of intelligent processing *per se*, but rather the imitation of human reactions, many of which are informed by our emotions, personal histories and social lives, and often have rather little to do with intelligence. Mimicking of all this so as to convince a discriminating judge over an extended period would be an extremely impressive programming achievement, no doubt. But the notion of intelligence – as Turing would have been the first to insist – is nothing like so human-focused as to require any such thing (cf. note 21 above), and hence it would be utterly perverse to make the general imitation of human characteristics a major focus of artificial intelligence research. Indeed this would be almost as ridiculous as making the imitation of birds – rather than fast, safe, and efficient flight – a primary aim of aeronautical research.[25]

7. Conclusion: The Turing Test and the Tutoring Test

When Turing proposed his "imitation game" in 1950, it served the valuable role of highlighting a context – namely textual conversation – in which one could realistically foresee computer behaviour that deserved to be called "intelligent". I have suggested that in this role the "Turing test" succeeds, and that if we were presented with a system which could reliably generate conversation of the quality he illustrates in his article, we would have excellent reason for counting it as intelligent, even though we would have no good reason for ascribing it any sort of conscious awareness. Admittedly this involves conceptual change, because our naïve concept of intelligence combines both *information processing* and *phenomenological* aspects, but such change is well motivated in this sort of situation, where we are presented with a new kind of entity that fails to fit into our naïve taxonomy. Moreover there are good independent reasons for seeing sophistication of information processing – rather than inner experience – as the central criterion for intelligence: this conforms to our standard methods of comparing intelligence amongst people and animals, and also acknowledges the reality of highly intelligent behaviour that is "intuitive", habitual, or subcognitive.

Unfortunately, however, the Turing test itself fares very badly as a method of *measuring* intelligence: it simply is not true that better performance in the test (in the sense of passing more plausibly for a human conversationalist, or for a longer period) correlates well with intelligent information processing. Nor is this only because success in the test is biased towards the imitation of *human* conversational behaviour, which surely disqualifies it as a *necessary* condition for intelligence (as Turing himself recognised). More damagingly, the development of chatterbots has revealed how unreliable we humans are as judges of conversational competence, mainly because we are so liable to read coherent meaning into any verbal exchange that is susceptible of it. Hence the chatterbot designer, aspiring to do as well as possible in the imitation game, aims not for the generation of precise and careful dialogue (in which the computer's

[24] The ascription could be withdrawn if it later turned out that the program was driven by a lookup table or just happened to "get lucky". As argued above, it is sophisticated and appropriate information processing – what enables it to respond "intelligently" to a wide range of inputs – that constitutes intelligence, not just outward behaviour on particular occasions.

[25] Whitby (1996), pp. 57-8 develops this point, while French (1990) highlights the extreme difficulty of programming a computer to mimic human *subcognitive* reactions.

mistakes or lack of "humanity" will become all too apparent), but instead for the production of piecemeal responses that are maximally vague and sloppy, exploiting the foibles of the interrogator. Thus there is no plausible developmental pathway from increasing chatterbot performance in the Turing test to genuine artificial intelligence, and the Loebner Prize (though no doubt well motivated) is completely misdirected.

It might, nevertheless, be possible to preserve something in a similar spirit if we add two letters and move from the "Turing test" to a "Tutoring test", in which the aim is not to pass for a human, but instead to succeed in a conversational information-processing task which has a very clear point and whose measurement is relatively well understood. Here the criterion of success would be not *deception* but *revelation*, by tutoring the human "interrogators" to acquire an understanding of some specific field of knowledge of which they were previously ignorant (e.g. some aspect of chemistry). In this context, for the tutoring system to reveal its non-human status would not be any kind of failure – all that matters is the effective eliciting of understanding in the tutee. And this can be assessed by the methods we standardly use in educational practice, ranging from first-personal reports to interviews and formal tests. Now the Turing-style gold standard would be a tutoring system that can teach as effectively (in a given time) as a good human tutor; and it is an open question, I believe, whether this is realistically achievable, and in which fields. But whether or not that provides a plausible ultimate target, the great advantage of this Tutoring test – in almost any field to which it might be applied – is that work towards it can potentially be of real value, not only in developing systems capable of providing cheap education to those unable to afford human tuition, but also in promoting *genuine* artificial intelligence.[26] For the comprehensive understanding of any intellectual issue by the tutee will typically involve the grasp of a complex web of connections amongst the relevant concepts and techniques. And to convey these most effectively, an intelligent tutoring system will presumably require some representation of these same connections: the more fully and faithfully they are represented, the better it is likely to be able to perform at tutoring. At any rate, it seems a relatively plausible expectation that work towards the Tutoring test could provide a valuable source of continuing inspiration in artificial intelligence. Sadly, the same can no longer be said of the Turing test.[27]

———————

This article is forthcoming in
Alan Turing: His Work and Impact,
edited by S. Barry Cooper and Jan van Leeuwen (Elsevier, 2012), pp. 39-52.

[26] Note the absence of any assumption here that "artificial intelligence" must be unitary: indeed the Tutoring test would suggest a domain-relative notion.

[27] For helpful comments on this paper, which have enabled me to improve it significantly, I am very grateful to Tim Bayne, Robin Le Poidevin, and Hsueh Qu.

References

Block, Ned (1981), "Psychologism and Behaviorism", *Philosophical Review* 90, pp. 5-43.

Dennett, Daniel (1995), "Intuition Pumps", in John Brockman (ed.), *The Third Culture: Beyond the Scientific Revolution*, New York: Simon & Schuster, pp. 182-8 (the rest of chapter 10, pp. 181-97, consists of commentary by others).

Floridi, Luciano (2008), "Artificial Intelligence's New Frontier: Artificial Companions and the Fourth Revolution", *Metaphilosophy* 39, pp. 651-5.

French, Robert M. (1990), "Subcognition and the Limits of the Turing Test", *Mind* 99, pp. 53-65 and reprinted in Millican and Clark (1996), pp. 11-26.

Hume, David (1748), *An Enquiry Concerning Human Understanding*, ed. Peter Millican, Oxford: Oxford University Press, 2007.

Michie, Donald (1993), "Turing's Test and Conscious Thought", *Artificial Intelligence* 60, pp. 1-22 and reprinted in Millican and Clark (1996), pp. 27-51.

Millican, Peter and Clark, Andy (1996), *Machines and Thought*, Oxford: Oxford University Press, 1996.

Millican, Peter (forthcoming), "Hume's Fork, and His Theory of Relations", forthcoming in *Philosophy and Phenomenological Research*.

Nagel, Thomas (1974), "What Is It Like to Be a Bat?", *Philosophical Review* 83, pp. 435-50.

Petzold, Charles (2008), *The Annotated Turing: A guided tour through Alan Turing's historic paper on computability and the Turing Machine*, Indianapolis: Wiley.

Preston, John (2002), "Introduction" to John Preston and Mark Bishop, eds, *Views into the Chinese Room*, Oxford: Clarendon Press, pp. 1-50.

Searle, John R. (1980), "Minds, Brains, and Programs", *Behavioral and Brain Sciences* 3, pp. 417-24.

Searle, John R. (1984), *Minds, Brains, and Science*, Cambridge, MA: Harvard University Press.

Searle, John R. (2002), "Twenty-One Years in the Chinese Room", in John Preston and Mark Bishop, eds, *Views into the Chinese Room*, Oxford: Clarendon Press, pp. 51-69.

Shanahan, Murray and Baars, Bernard (2005), "Applying global workspace theory to the frame problem", *Cognition* 98, pp. 157-76.

Sober, Elliott (2000), "Evolution and the Problem of Other Minds", *Journal of Philosophy* 97, pp. 365-86.

Turing, A. M. (1936), "On Computable Numbers, with an Application to the *Entscheidungsproblem*", *Proceedings of the London Mathematical Society*, Second Series, Vol. 42 (1936-7), pp. 230-65.

Turing, A. M. (1950), "Computing Machinery and Intelligence", *Mind* 59, pp. 433-60.

Weizenbaum, Joseph (1966), "ELIZA – A Computer Program For the Study of Natural Language Communication Between Man And Machine", *Communications of the ACM* 9, pp. 36-45.

Whitby, Blay (1996), "The Turing Test: AI's Biggest Blind Alley?", in Millican and Clark (1996), pp. 53-62.

Williamson, Timothy (1994), *Vagueness*, London: Routledge.