Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Connecting the dots my own way: Sphex-test and flexibility in artificial cognitive agents.

Juan Camilo Espejo-Serna[1]

Abstract. The purpose of this paper is to raise doubts on the standard approach in cognitive science regarding the design of artificial intelligence. An intelligence criterion, different from the Turing test, is presented and explained: the sphex-test. Following the explanation of the test, a theoretical approach of the normative constraints of experimental setups is presented and the causes of failure determined through it. It is shown that domains with which the agent can only relate in one unique way (have only one presentation mode) are responsible for failure to pass the test. Finally, a possible solution to the problem is suggested by sketching one way of obtaining the necessary flexibility to not to fail the sphex-test.

From now on, I'll connect the dots my **own** way.
Calvin, from *Calvin & Hobbes*

Cognitive science attempts to give an explanation of what intelligence is and how organisms attain it. At first sight, the design of an artificial agent able to pass the Turing test could give us the key of this enterprise. The goal of the Turing test is to offer a criterion to determine if machines can think. Thus, an artificial agent that passes the Turing test would be legitimately attributed cognitive abilities, *i.e.*, we would attribute intelligence to it. Even if the Turing test is a good criterion for the attribution of intelligence, it only presents a sufficient condition for intelligence: an agent could fail to pass the Turing test and still have some kind of intelligence. In this sense, the Turing test is too strong: the failure to pass the test cannot guarantee the lack of cognitive abilities whatsoever. We need a test to determine intelligence such that, failure to pass it consists a proof of the lack of intelligence, *i.e.*, a necessary condition for passing the test is the possession of intelligence.

The question of what intelligence is and how organisms attain it can be reformulated as follows: what is the simplest activity system which is such that it requires intelligence to perform well at it? The standard assumption in AI in order to deal with this question is to pick a domain for which cognitive abilities are needed to perform adequately and design an artificial agent that performs adequately in such domain, and from there determine whether the artificial agent has certain cognitive abilities, and whether its architecture provides an explanation of how such cognitive abilities work.[2] In this paper I want to question this standard assumption. I will present a simple test —the sphex-test— that any artificial agent under the described approach should pass, and apply it to a paradigmatic project in the new artificial intelligence: Rodney Brooks' *creatures* (2; 3). Based on the sphex-test, I will suggest that flexibility is a necessary condition for the attribution of cognitive abilities to artificial agents, and so that the evaluation of an agent performance in a domain is not enough. I will then focus on finding a common denominator that can explain failures to pass the test, and argue that in order to ascribe cognitive abilities to an agent it is necessary to consider the presentation mode of the domain it is performing in. The complexity of proper performance in a domain changes according to the way the domain is presented. Different presentation modes ask for different behaviors of different complexity; such that these three elements (the domain, the performance and the presentation mode) are necessary for proper cognitive abilities ascription.

A general characterization of some normative constraints (bounding, constitutive and governing) is used to pin point the deficiencies and detect a particular kind of domain guilty of them. The problem is found in cases where there is a TICTACness domain: a case when there is no space for difference between the bounding normativity and the constitutive normativity. Once the problem is detected, a possible way out is presented: a domain where the presentation mode from which the agent starts has no direct relation with the correct performance in the domain. Artificial agents should behave as Calvin does: when presented with a domain, they should built up their own way of dealing with it, not restraining themselves to one way the domain is presented. Cognition is not seen when the agents perform just as the domain is initially presented: by simply connecting the dots in the predetermined manner. Artificial agents must be able to connect the dots on their own, thus truly exhibiting cognition.

1 Sphex-test and flexibility

I want to propose a simple test for any particular cognitive science project that attempts to show something about cognition through the behavior of artificial agents. I wish to test the correctness of such inferences. The idea is to *slightly* modify the environment in which the agent performs and examine if the original claims made about the agent still hold. A change in the environment should only cause a modification of the agent's performance (*v.g.*, render it not as well-adjusted) but should not modify the kind of behavior, *i.e.*, it has to be possible to describe the agent's behavior in the same way albeit with small qualifications. For instance, if an artificial agent is described by its designers as exemplifying a plan-following behavior inside a given environment, it should also be possible to describe that same agent as following plans in the modified environment (but perhaps in a sloppier manner).

Let us take the example of the digger wasp (wasps of the genus *Sphex*). Inside its niche the digger wasp can be said to have a plan-following behavior: it "wants" to take care of its eggs and has a plan

[1] National University of Colombia. Email: jcespejos@unal.edu.co

[2] Exactly this is Shannon's argument for attempting to build a computer that can play chess like a master: if playing chess like a master requires intelligence and a computer manages such skill of game play, then such computer would be intelligent, see (9).

for it. The wasp looks for a prey, stuns it, digs a burrow, hides the prey inside and lays its eggs besides it so that the newly born larvae will have food right away. If the original description of the wasp's behavior as plan-following was correct, then an alteration in the environment would not change it radically. (Although —we would imagine— the wasp would have to modify its original plan in order to cope with the changes). This situation is presented by Woodridge (11), where the wasp's catch-dig-bury procedure is interrupted by an experimenter that slightly changes the position of the stunned prey while the wasp is checking that the dug burrow has nothing unexpected (another predator ready to eat both eggs and prey, for example).[3] If the description of the wasp's behavior as plan-following was correct, then in this modified environment the wasp would still develop a plan. Yet the wasp cannot handle the modification. As the wasp goes inside the burrow and the prey is moved by the experimenter, and then the wasp comes out, it returns the prey to its original position and reexamines the burrow. This is repeated *each* time the prey is moved. The experimenters moved the prey up to forty times and the wasp kept on repeating the same action all over again. Not a single time did the wasp ignore the change and move on. The initial cognitive explanation for the behavior of the wasp was proven incorrect by a simple modification: the allegedly plan-following behavior of the digger wasp was revealed as absurd repetition.

The wasp, that in one scenario behaved as if it had cognitive abilities, behaves as lacking such abilities in the modified scenario: in one case behaved as if it could follow plans, while in the modified case it behaves as repeating an absurd routine. If we are going to ascribe cognitive abilities to an agent, it —unlike the wasp— should be able to perform similarly in a different environment. If an agent really has the purported cognitive abilities it should be able to exercise such abilities in more than a limited context. Based upon this we can construct a test for artificial agent: if an artificial agent fails it, the cognitive theory sustained by it would be proven incorrect. Through this test we will be able to distinguish cognitive descriptions *proper* of an agent from cognitive descriptions *read into* an agent. A cognitive description proper of an agent will survive the sphex-test, while cognitive descriptions read into the agent will fail to pass the test. This will be called the sphex-test:

> ***Sphex-test***: If it is possible to introduce a small modification in the environment that causes a massive collapse in the appearance of the alleged cognitive abilities, then such abilities are only read into the agent.[4]

In order to better understand the basics of the test let us apply it to a paradigmatic example of the new AI: Rodney Brooks' *creatures* (2; 3). These robots are designed to perform fluently in a natural environment: not a controlled experimental setup but rather a real world environment. The subsumption architecture behind them allows layers with different functions to work in parallel, such that the creatures cope appropriately and in a timely fashion with changes in its dynamic environment. Minor changes in the properties of the world should not lead to total collapse of their behavior, they can both adapt to surroundings and capitalize on fortuitous circumstances and have some purpose in being (4, 4). One of his better known robots is Herbert, designed to pick up empty soda cans around the MIT labs, from which he says has the cognitive ability to move around picking up empty cans. It has obstacle avoidance, wall following, object recognition (empty soda cans and desks) and other behaviors that allow it to search, reach and pick up soda cans (3). Herbert is designed to function properly inside the MIT lab, where the doors, walls, desks, *etc.*, look a certain way. Although perhaps not in an explicit way, Herbert is designed to move around the kind of objects encountered in the MIT lab, those that during debugging forced adjustments. If Herbert was transfered to another lab, its performance would probably lower or fail because its basic layers are not adjusted to the new environment. For example, if the MIT starts buying different sodas (with bigger and heavier cans) Herbert would leave them alone regarding them as not empty. (If the new empty soda cans weight the same as the old cans that still have some soda, Herbert will not pick up those cans.) If the sphex-test is applied to Herbert at this point, it would fail: a small change in the world (the change of weight of the cans) would make it fail in its task (pick up the empty cans).

One might say that this failure does not make Herbert collapse altogether because Brooks' crew could redesign it, as part of the debugging procedures, allowing it to handle the new kind of cans. If the cans are not being detected by the sensory apparatus, they will be modified in order to allow Herbert to behave suitably again. But such changes would be *ad hoc* and offer no guarantee of flexibility for future cases. Even more, if an external agent is the one that makes the adjustments, then a second order sphex-test can be formulated. Can the agent that modifies Herbert respond flexibly to changes in the conditions to which it must adjust Herbert? If the answer to this question is negative, then Herbert cannot pass the sphex-test because there might be a situation for which both Herbert and the engineer are not prepared. But if the answer is positive, the mechanism that accounts for this flexibility will also be responsible for Herbert's flexibility, and thus the cognitive abilities should be ascribed to such mechanism and not to Herbert. Eitherway, there are no grounds of the ascription of cognitive abilities to Herbert. If there is no way to guarantee flexibility, the correctness of the artificial intelligence enterprise is undermined. Without flexibility there is no way to be sure that the cognitive abilities the designers want to attribute to the machine are proper or just read into it and thus no legitimate way of extracting cognitive conclusions from physical models. Without an answer, there are only mechanichal systems with rather complicated behaviors. It seems then that flexibility is the mark of cognition.

The sphex-test is not meant to prove that a given agent has no cognition whatsoever. Someone could argue that, in a sense, even an ordinary thermostat has at least some kind of cognition: it has an access to the world and produces changes accordingly. This implies that to conclusively determine that an agent has no cognition asks for a detailed examination of what cognition is and how can it be exemplified: the necessary conditions of any kind of cognition should be examined. The sphex-test has a more modest reach. It aims to determine the correctness of cognitive abilities *ascriptions* to artificial agents based on a particular behavior. The digger wasp could be an intelligent agent wishing to conceal its abilities by acting stupidly, but the behavior exhibited cannot justify the ascription of such abilities. Likewise, an artificial agent that fails the sphex-test could have some kind of cognition (at least in the weak sense of the thermostat) but lacks that that was supposed to have. Cognitive scientists attempt to justify cognitive abilities ascriptions to an artificial agent because it behaves in a certain manner in a domain. The sphex-test undermines the possibility of such inference by presenting cases in which

[3] The idea of using the case of the digger wasp to examplain cases of failures or breakdowns in intelligence is from Adrian Cussins. For example, see (5).

[4] The reader is asked to accept for now this intuitive understanding of the sphex-test, in particular what regards the conditions for a change in the domain to be a "small modification". The small change must be such that the agent can still perform (perhaps badly) whatever it was doing before the change. The characterization offered in the following section should help settle possible doubts regarding particularities of the test.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

an agent that has proper performance in a domain fails to exhibit intelligence when faced with a small change, thus suggesting that there is more than these two aspects involved, *i.e.*, other elements besides the domain and the behavior are at issue.

2 Appropriate domains

As we have seen in the previous section, there are reasons to doubt about the possibility of obtaining conclusions from the behavior of artificial agents in domains. The sphex-test suggests that there is another factor in play here, that not just any task and not just any behavior can have explanatory impact and elicit conclusions about the nature of cognition. This asks for a general characterization of which domains and which behaviors do not have explanatory impact and why, and which kind of domain and behavior, if any, could have such impact. In this section it will be claimed that a TICTACness constitution of a task domain is the reason of failure to pass the sphex-test, and so artificial agents must not have TICTACness domains. In order to do so, a general way to characterize tasks and behaviors in terms of their normative constraints will be presented and through this pin point a particular kind of constitution —TICTACness— that is common to all projects that fail to pass the sphex-test.

2.1 Domain, representation and valuation

The standard approach in cognitive science is to construct an artificial agent that performs adequately in a domain for which certain cognitive abilities are necessary. The domain, then, has to be such that cognitive abilities (and not just causally explained relations) are needed for proper performance, *i.e.*, the domain and the performance have to be a certain way. These aspects (the domain and the kind of performance required) directly determine if cognitive abilities are needed, and thus if an artificial agent performing adequately has them.

Think about tic-tac-toe (TTT). TTT is a two-player game generally played in a nine slot board, where players take turns placing different tokens in it. There are three possible board states that end the game: when the first player's tokens are in a δ combination, when the second player's tokens are in a δ combination, or when there are no more empty slots in the board. These rules and an explicit characterization of what a δ combination is delimit the *domain* of TTT, and consitute the minimum amount of information a player without any previous particular knowledge about the game must have in order to start playing. Yet this information can be *represented* in a manyfold of ways (have different modes of presentation).[5]

There is an intricate connection between the representation and the possibility of achieving a δ combination. For a regular adult, playing TTT in a 3×3 grid would be much more easier than playing TTT using another presentation of the domain[6]. Thus, the particular way of representing the domain has a role in the determination of the kind of performance within the domain. The domain can remain the same but a change in the way it is represented could change the transformations needed for proper performance.

It is not enough to examine the performance in a domain in order to ascribe cognitive abilities to an artificial agent. The particular way the domain is presented (the representation the agent has of the domain) plays a major role. Projects in cognitive science must be examined regarding the domain the agent performs, how the agent performs *and* the representation of the domain used to achieve such performance. These three factors will determine the validity of cognitive abilities ascriptions.

Remember the case of Herbert. The *domain* Herbert is designed to work in is the MIT labs. If Herbert is going to do any task (correctly or incorrectly), it must be able to comply to the general characteristics of the MIT labs. One way this domain could be presented would be in terms of objects and their properties, *i.e.*, in terms of walls, corridors, people moving around, laptops, sodas, *etc.*. Achieving the goal of picking up empty soda cans in terms of this would require having a model of the MIT labs such that a trajectory evading obstacles and recognizing and picking up cans is made possible. Yet this would need the computation of all the changes in the environment that could produce a modification in the model of the world, adjusting the model to them, and then deciding a new trajectory for the current state of the world. A feat of huge computation powers. But Herbert does not behave in this way. It deals with its domains directly as it is presented by its sensory channels, without any further processing or modelling. The domain is presented only as sensory data that causes particular movements. The subsumption architecture of layers deals with the different stimuli producing a combination of movements that result in a particular behavior that fulfils its goal. As in the TTT case, there is a way of distinguishing three determinations: Herbert's domain (the MIT labs), the way the domain is presented to it (sensory information) and its performance.

These examples present two different phenomena characterized in terms of the boundaries of a domain, its presentation and what proper performance is. This difference can be characterized as the difference between aspects that *delimit* the domain, aspects that *constitute* the domain and the *valuations* of the elements in the domain. We will call these normative aspects bounding normativities, constitutive normativities and governing normativities, respectively. The *bounding normativity* delimits the domain: what belongs and what does not to the domain; it is the one in charge of delimiting the space of the domain. The *constitutive normativity* characterizes the elements of the domain: the way the domain is presented. The same domain (the labs) can be presented in terms of objects and their properties or as stimuli and this normative constraint regulate which way and how the domain is presented. And lastly the *governing normativity* is the standard against which it is determined what correct performance in the domain is (closeness to a goal).[7]

2.2 TICTACness and failures to pass the sphex-test

Let us take now the understanding of TTT in terms of the three kinds of normativities and explain why it is prone to produce inflexibility cases. Imagine a computer program that plays TTT like a master (it plays just like any regular adult would play).[8] Yet, if this is true and

[5] For example, it can be presented as a game played in a 3×3 grid where the δ combination is made by placing three in a line or like a game played with nine different chips where the δ combination is made by a particular combination of chips. What is important is that there is a method to translate one way of presenting the game to the other, thus warrant the sameness of the game.

[6] For example on in which there are nine chips labeled with a different element element x, such that $x \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$, and where the δ combination for this case occurs when an element d, such that $d \in \{7, 56, 73, 84, 146, 273, 292, 448\}$, can be expressed as the sum of some of the numbers selected by a player.

[7] This understanding derives from the normative characterization presented by Adrian Cussins in several conferences and seminars, with slight differences rooted in my understanding of similar remarks presented by John Haugeland (6).

[8] Heavy programming skills are not needed to design a simple program that can perform like a master in TTT. A fast google-search for TTT

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

playing TTT requires cognitive abilities to play, then, according to the assumption of cognitive science, such program must have some cognitive abilities, namely those that master human TTT players also have. But a TTT-playing computer can be excellent at playing simple TTT, and collapse in an TTT-variation[9]; and thus fail to pass the sphex-test. That is, a TTT-playing computer is prone to fail the sphex-test.

Why does the TTT-playing computer fail the sphex-test? It is because (a) the specification of the TTT domain is necessary in the design of an artificial agent and (b) such specification itself is enough to prescribe one unique representation and determine how appropriate behavior in the domain should be. If a and b hold, then (c) what is considered appropriate behavior for the program is completely fixed. But if c, a small change in the environment (such that produces a change in the valuation) could produce a massive collapse (*v.g.*, in the case that what was once incorrect behavior is now proper). Just like years of natural selection have determined the adequate behavior for the digger wasp, the designers determine what the adequate behavior of the TTT-program should be. And in the same way, the TTT-program is subject to the problems of the wasp, *i.e.*, fail the sphex-test.

It is obvious that an artificial agent cannot start completely out from zero. The designer has to determine *some* of its characteristics. Since the least amount of information needed to start to play TTT are the rules, the domain must already be fixed in the programming of the artificial agent. This means that thesis a cannot be rejected.[10] Then, what makes a TTT-playing program fail the sphex-test is that the domain of TTT itself imposes the constitutive and governant normativities (thesis b). That is, the determination of the domain (the rules of TTT) dictates what the elements of such domain are (the representation) and a unique method to achieve the goal. Accordingly we can say, in general, that what produces inflexibility in a given domain is that the constitutive and/or governant normativities get fixed by the domain. I will call this a TICTACness scenario: a case when the bounding normativity fixes the constituent normativity and/or the governant normativity. In the case of TTT the domain itself limits the representations to an equivalent set, and also how to produce δ conditions (how to win),*i.e.*, they determine a unique procedure to cope with the elements in a way that the goal is achieved.

Reconsider the case of Herbert and the sphex-test scenario formulated for it: a modification of the weight in the cans produces a massive collapse. A change in the weight of the cans produced a collapse of the appearance of cognitive abilities —Herbert completely ignores the goal and fails the sphex-test. The formulation of the sphex-test makes use of a small modification in the domain that cannot be accounted by Herbert's sensory apparatus (the representation of the domain[11]). Herbert can perform adequately in its domain because its inputs are associated with movements in a way that results in the fullfiling of its goal. Brooks' crew tweaks these associations in a way to guarantee it. But if there is an aspect of the domain for which there is no associated response, Herbert will not be able to deal with it. Precisely this is what happens in the sphex-test scenario: a new kind of input (the new empty can weight) for which the old procedures are not linked. This shows that the representations of the domain (Herbert's input-output associations) directly determine the possibility of success in the domain. In other words, Herbert exemplifies a case of TICTACness. There is only one way the domain is presented, thus being equivalent to the domain: whatever the representation cannot account is excluded from the domain. This leaves open the possibility for a sphex-test failure: if there is a change within the domain there is no way that Herbert can notice it, behaving exactly as if nothing had changed; with the result of ignoring the empty soda cans and thereby failing the sphex-test.

algorithms presents tons of possible solutions that can be easily reproduced.

9 Consider a tic-tac-toe game where the goal is to make the opponent make three in a line, or any other novel standards for achieving victory.

10 One could think that through methods like genetic algorithms a specification might be achieved that is not explicitly made by the programmer. Yet the method itself is determined (how a genetic algorithm works), so in a way the results are also determined, just in a broader way.

11 The representation of the domain is by no means a symbolic representation. The representation Herbert has is just the way that the world is presented to it: a bundle of information received through its receptors.

3 Achieving flexibility

We can now reexamine the intuitive formulation of the sphex-test we started from, making use of the characterization in terms of its normative constraints, in a way such that a possible solution begins to be visible. This will help us give a better explanation of what a small change is. A change, in general, is a modification of the initial situation of the artificial agent. And since the initial situation is characterized in terms of its normativities, a change in the situation is then a change with respect to the bounding, the constitutive or the governing normativity. Accordingly, a small change is one that (a) is not a change in the domain (a change in a domain would be a major change), thus falling within the scope of the bounding normativity and yet (b) does not comply to the constitutive and governing normativities. But if b, then such element is prone not to produce proper performance since the governing normativity cannot give a value to its contribution in reaching the goal. Thus, we can have a new formulation of the sphex-test:

> ***Sphex-test revisited***: If there is an element that falls within the bounding normativity for which there is not an element within the constitutive normativity that conforms with the governing, the alleged cognitive abilities correctly are only read into the agent.

This formulation presents a possible way out: in order to achieve the necessary flexibility to pass the test the presentation mode of the domain built in the agent must not have a predetermined connection with the steps necessary to achieve the goal the achieving of which will determine the grounds for cognitive abilities ascription. Such connection has to be created by the agent based upon some other mode of presentation that has a different goal and which has no evident relation with the first goal. This hypothesis tries to capture the fact, showed around the presented example, that cognition is involved both in the the act of *dealing* with a representation and *achieving* a certain representation. This is a fact ignored by several approaches in cognitive science: projects inscribed in the so-called good old fashioned artificial intelligence (SHRLDU(10), CYC(7), *Deep Blue*), others that make use of allegedly deictic representational systems (Pengi (1)), subsumption architectures (Herbert) or neural networks (NETtalk (8)), to name just a few. There is a failure to note the fact that the very same act of having a particular representation involves cognition, precisely the kind of cognition marvelously present in the artificial agents. There has to be a difference between the way the domain is presented and the way the domain is, such that being excluded from the representation does not necessary imply an exclusion from the domain. In other words, there must be space for error and change in the mode of presentation of the domain.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

3.1 Possible ways to find a solution

One would imagine that stating the possibility of different goals within one same domain would be enough to dissolve the possibility of TICTACness. Let us examine two particular cases of a bounding normativity where two different governing normativities can apply.[12] One example of this can be seen in the relation between TTT and TTT^{-1}. TTT^{-1} (inverse TTT) is exactly like TTT except that the δ conditions for the players switch between them: one player wins when the *other* makes three in a line. This means that the domain for both games is exactly the same while their goals change. But as we saw above, the bounding normativity of the TTT domain fixes a unique set of equivalent representations; thus, if TTT^{-1} and TTT have the same domain, their constitutive normativities are also the same. A different example can be seen between *Atari-go* and regular *Go*. *Go* and *Atari-go* are oriental games played by two players who place black or white stones, respectively, anywhere on the intersections of a 19×19 grid, as long as the stone placed doe not recreates a former board state. Stones remain in the position they were placed unless they get surrounded by enemy stones, thus captured and consequently removed from the board. The goal of *Go* is to control the largest portion of the board (territory), while in *Atari-go* the goal is to capture the largest amount of enemy stones (regardless of the amount of territory controlled at the end of the game).

In the case of TTT and TTT^{-1} the goals, although different, are mutually inter-definable: one can understand TTT^{-1}'s goal in terms of TTT's goal. But this is not the case for *Atari-go* and *Go*: having the largest amount of territory does not necessarily amount to having captured the largest amount of stones and, likewise, capturing the largest amount of stones does not guarantee having the largest amount of territory. Since we want to avoid the cases where the representation alone assures compliance with the governing normativity, it is important that domains that can only have one set of mutually defineable goals are not selected. So we can say that a good sign that a domain is TICTACness is the impossibility to state different non-interdinable goals for which cognitive abilities that require cognitive abilities for proper performance. Yet having the possibility of having different goals does not guarantee flexibility, at least not on its own.

The sphex-test scenario presented shows how the possibility of different presentation modes of the domain is completely ignored, thus spoiling the appearance of cognitive abilities. Such failures are produced because the kind of performance expected is directly determined by the built in representation, product of their TICTACness. For instance, the behavior of the digger wasp in its original environment is the best course of action, but the inclusion of a small change (the movement of the prey) makes a change within the domain for which there is no adequate representation, such that what was the best course of action —what gave the appearance of intelligence— shows itself as absurd repetition (reveals the folly of the agent).

Maybe for natural agent, as the wasp, it is not easy to distinguish what falls into its domain and what does not. The evolutionary history of the organism and its niche may be not fully understood and so provide no easy way to draw the fine line between what is within its domain and what is not. However experimental setups are, or at least attempt to be, controlled setups where the task can be effectively done. Even in the case of Herbert where the domain are the MIT labs (not a simple fake setup like SHRLDU's, claims (3)) there are aspects excluded from the possibilities within the domain (for example, rivers insteas of corridors). The *quid* of the sphex-test is to present an example of a small change within the domain for which the agent has no adequate response but should have one. This means that in order to answer the challenge of the sphex-test it is necessary for the agent to have a way of dealing with the domain and to cope with unexpected elements (elements that are not initially represented). That is, that whatever presentation mode of the domain programmed (the representation the agent starts from) it must be possible to have other non-equivalent representations for that same domain. The sphex-test scenario are cases of elements that fall within the domain but do not fall within the representation the agent has of the domain. Thus, the way to solve this problem is for the agent to be able to handle, within the domain, different complete non-equivalent nor translateable representations. Such presentation modes of the domain have to be such that they are still presenting the *same* domain yet in a different manner.[13]

I will try to explain myself through yet another example. Consider drawing. The goal of this form of art, one could say, is to produce drawings, certain figures that resemble objects in the world. This simple form of art has clear definite bounding normativity that excludes other activities related with paper and pencil (*v.g.*, writting). Yet the governing normativity is not yet specified: nothing is said about the kind of figures and the methods that help achieve them. It is important to distinguish these two normativities: one determines the boundaries of the domain (what makes it drawing and not writting) while the other determines the way the domain is presented. Now take pencil and paper and draw a duck. The statement of this chore will only tell you the starting point (a blank sheet of paper) and end conditions (sheet of paper with a drawn duck). *How* to transform the blank paper into a drawing of a duck is yet to be determined: there is no fixed method to relate the final outcome to a sequence of pencil strokes starting from the empty paper. Having a paper and pencil and the goal of drawing a duck places you within the domain of drawing (excluding others like soccer, writting or reading). Consider now that you have a connect-the-dots diagram for the duck: it might require some skill but there is a clear method to relate the elements to the end (connecting with a line consecutive numbered dots). The dots determine which pencil strokes are correct and which ones not: they consitute one unique path to transform an empty sheet of paper into a drawing of a duck. A person may be very skillful in drawing a particular figure (following a diagram) and completely fail to build another figure (for which she has no diagram). If the only grounds for attributing cognitive abilities to that person was her ability to draw the duck, it would fail the sphex-test, leaving the ascription without grounds. This means that whatever elements an artificial agent starts from they must not be such that the main goal can be directly understood through them. Otherwise the cognitive workload is carried by the predetermined representation, and thus by the designer that predetermined such representation.

4 Conclusions

The situation of artificial agents is like Calvin's one in the comic strip from where the quote at the beginning in the introduction. In the strip, Calvin is complaining to Hobbes that he was forced to draw a duck by the connect-the-dots diagram. The point is that if such diagram is followed blindly, the result is fixed and thus inflexible, producing the

[12] It might be the case that the governing normativities be of the same kind (the same kind of guiding) but having different goals is enough to distinguish them.

[13] In order to continue the path presented in this work it is needed a full account of how is it possible to have different complete presentation modes of the same domain with different semantic values. Spelling out the full implications of this approach is out of the scope of this paper.

duck Calvin despises. Likewise, if the presentation mode of the domain is followed blindly (it is the only guide), the result is fixed and inflexible: TICTACness. The solution for both cases is to be able to use the initial presentation (the dots in Calvin's case) and construct something not directly specified: a novel drawing (product of connecting the dots in his own way) or a novel way of performing. Such novel constructions, ones that are not hidden within the original representation but constructed from them, are the ones that exhibit true cognition.

The sphex-test attempts to show that flexibility is needed for proper attribution of cognitive abilities. This flexibility is not a matter of adaptability, but of the normative structure of the experimental setup: the domain, the performance and the way the performance is achieved are crucial in determining if a given artificial cognitive agent is flexible or not. Accordingly, to succeed in the search of artificial intelligence it is not so important the complexity of the behavior but of the qualities of such behavior. For example, *Deep Blue*'s behavior can be highly complex but utterly useless in broadening our understanding of what intelligence is. If an artificial agent is to advance our understanding of cognition, it is necessary that it performs in a special domain, in a particular manner and with respect to a certain task. Thus, it seem that only if an artificial agent can pass the sphex-test can we hope to achieve artificial intelligence.

ACKNOWLEDGEMENTS

This work has been made within the *Perception, Coordination and Communication* research group of the Philosophy Department at the National University of Colombia. I would like to thank its members for fruitfull conversations around the subject. I would also like to thank Adrian Cussins, Carlos Márquez, Diana Acosta, Felipe Cuervo and an anonymous referee for their valuable comments on previous versions of this paper.

References

[1] Philip E. Agre and David Chapman, 'Pengi: an implementation of a theory of activity', in *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, pp. 268–272, (1987).

[2] Rodney A. Brooks, 'A robust layered control system for a mobile robot', *IEEE Journal of Robotics and Automation*, **2**(1), 14–23, (1986).

[3] Rodney A. Brooks, 'How to build complete creatures rather than isolated cognitive simulators', in *Architectures for Intelligence*, pp. 225–239. Erlbaum, (1991).

[4] Rodney A. Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**, 139–159, (1991).

[5] Adrian Cussins, 'Why a closed, rule-governed, digital system need not be a formal system: an argument concerning the ancient game of go, the nature of computation and the distinction between intelligent creatures and mere brutes', (1999).

[6] John Haugeland, *Having Thought. Essays in the Metaphysics of Mind*, Harvard University Press, Cambridge, Massachusetts, 1998.

[7] Douglas B. Lenat and R. V. Guha, *Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project*, Addison-Wesley, 1990.

[8] T. J. Sejnowski and C. R. Rosenberg, 'Parallel networks that learn to pronounce english text', *Journal of Complex Systems*, **1(1)**, 145–168, (1987).

[9] Claude E. Shannon, 'Programming a computer for playing chess', 2–13, (1988).

[10] Terry Winograd, 'A procedural model of language understanding', in *Computer Models of Thought and Language*, eds., Roger Shank and Kenneth Colby, W. H, Freeman, (1973).

[11] Dean Woodridge, *The Machinery of the Brain*, McGraw Hill, Ney York, 1963.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Causal and Communal Factors in a Comprehensive Test of Intelligence

Paul Schweizer[1]

Abstract. The paper begins by examining the original 'intelligence test' as proposed by Turing, the discipline of 'Strong' Artificial Intelligence that followed on from Turing's work, and Searle's high profile critique of them both. In tracing the ensuing dialectic between Searle and his own critics, I support Searle's rejection of the 'Systems Reply', and for reasons based more on the philosophical views of Putnam agree that the original Turing Test (TT) is essentially inadequate. The situation becomes more complex with the 'Robot Reply' and allied Total Turing Test (TTT), and I argue that Searle's attempted refutation of the combined Robotic-Systems reply is unconvincing. However, this is not to say that the position expressed in this reply is itself confirmed, and I argue that externalist views in the philosophy of language first put forward by Kripke and Putnam cast serious doubt on the issue.

In turn, the causal and communal factors highlighted by externalist views in the philosophy of language point to the need for a fundamental shift in conceptual perspective and a strengthening of criteria in a truly Comprehensive Intelligence Test (CIT). I argue that an ideal CIT should focus on the *category* of cognitive system as a whole, rather than on the performance of individual artefacts. From this expanded perspective, the central question is not whether an isolated agent could simulate human performance within the context of a pre-existing sociolinguistic culture developed by the *human* cognitive type. Instead the key issue is whether the artificial cognitive type *itself* is capable of producing a comparable sociolinguistic medium of intelligence, where this essential medium is simply taken for granted as a precondition of the individual performances evaluated in the TT and TTT.

1 THE TURING TEST AND 'STRONG' AI

What would be required for a computational artefact to count as genuinely intelligent in manner comparable to a human being? In 1950 Alan Turing [1] famously proposed an answer to this question, and the controversy launched by his position is still underway. Turing replaced his opening question 'Can (or could) a machine think?' with the more precise and empirically tractable question 'Can (or could) a machine pass a certain type of test?', where the test criteria are framed in terms of *behaviour* that is standardly held to signify intelligence in the case of humans beings. In particular, the original 'Turing Test' (TT) is based entirely on *linguistic* inputs and outputs, and is designed as a free ranging session of questions followed by anonymous verbal responses. Linguistic performance is an apt choice as a pivotal criterion of intelligence, since human language is perhaps our most distinctive feature as cognitive agents, and it is an essential medium through which most of our higher level mental achievements are developed and expressed. Hence human language will retain a central role throughout the ensuing discussion.

The TT itself is a disputed issue and many of Turing's successors in the field of Artificial Intelligence would independently endorse a basic computational theory of the mind, wherein mentality is held to be explained via the physical realization of the right type of abstract computational procedure, without accepting the TT as an adequate criterion for deeming that a machine possesses genuine intelligence. The two issues are clearly separable, and one can embrace a computational approach to the mind without accepting Turing's original and quite controversial standard. According to the wider project embraced by 'Strong' AI and the computational paradigm, the relation between the abstract program level and its realization in physical hardware could still give an answer to the problematic nature of the relation between mind and body: the mind is to the brain as a program is to the hardware of a digital computer. On this model, computation could then be seen to provide the scientific key for explaining mentality and intelligence. Cognition in general (including the human case) is to be literally described and understood in computational terms.

2 SEARLE'S BASIC CRITIQUE

Probably the most high profile criticism, both of the TT in particular and the computational paradigm in general, was provided by the philosopher John Searle in his 1980 paper 'Minds, Brains and Programs' [2], where Searle put forward his celebrated Chinese Room Argument (CRA), designed to refute the view that passing the TT is a sufficient condition for genuine understanding or intelligence, as well as the more wide-ranging position that a system can sustain genuine mental states in virtue of instantiating the right type of abstract computational procedure. According to Searle, computation per se is neither a necessary nor a sufficient condition for the presence of mental states and real intelligence, and so he deems computation to be fundamentally irrelevant to the issue. Hence both the original TT and the more general mind/program analogy are summarily rejected.

As is well known, the CRA is based on a thought experiment in which Searle, a native speaker of English who knows no Chinese, is locked in a room with a massive rule-book written in English. He receives Chinese inputs on bits of paper and mechanically follows the instruction manual to produce outputs in Chinese script. For the sake of argument, we are asked to suppose that the manual is so good that he is able to fool native speakers of Chinese. They are outside the room giving him the inputs which are questions in Chinese, and by following the recipes for symbol manipulation provided in his rule book, Searle

[1] School of Informatics, Univ. of Edinburgh, EH8 9AD, UK. Email: paul@inf.ed.ac.uk.

is able to produce appropriate Chinese responses, and on the basis of these outputs, his questioners conclude that a fluent Chinese speaker is locked inside the room.

But Searle doesn't understand Chinese, doesn't even know basic Chinese vocabulary, and has absolutely no idea what he's been asked or what he has 'answered'. Yet he has passed a Chinese version of the Turing Test by serving as a human implementation of the conjectured program, and he has done so purely on the basis of formal transformations performed on 'meaningless' syntax. Searle's broad sweeping conclusion: mere success at the 'imitation game' and passing the TT is theoretically inadequate as a criterion of intelligence, and instantiating a computer program has nothing to do with genuine understanding or mentality. But in response to Turing's original question 'can a machine think?' Searle says the answer is definitely y*es* – because *we* are biological machines, and we can think. According to him, we can think in virtue of the unique physical structure and real causal powers of the actual brain – not in virtue of some abstract *formal shadow*, either classical or connectionist.

3 THE SYSTEMS REPLY

Searle addresses some anticipated objections to his view, the first such he dubs the 'Systems Reply'. A defender of the computational paradigm might argue that perhaps Searle in isolation doesn't understand Chinese, but that's not the point, because the *whole system* that produces the behavior – room plus manual plus Searle – does understand Chinese. Searle responds by claiming that *he* is the only locus of understanding in the scenario, and if he doesn't understand Chinese, then nothing else about the system does. It's absurd to imagine that even though he himself doesn't understand Chinese, somehow the conjunction of Searle plus pencil, slips of paper, instruction manual, four walls, etc. understands Chinese. His pencil and the four walls don't understand *anything at all*. Furthermore, suppose that Searle were gifted with a photographic memory, and could memorize the rule book. Then the entire set-up could be internalized, and Searle could perform the rule governed manipulations sitting outside under a tree. Searle himself would then *be* the system but he still wouldn't understand Chinese.

I would agree with Searle that the System's reply is not an adequate rejoinder, mainly because (i) the rest of the system in the original CRA scenario isn't *doing* the right sort of thing and (ii) the standard TT is woefully inadequate in any case (but not because of the introspective angle propounded by Searle – more on this later). At this juncture I would concur with Hilary Putnam's [3] basic critique of the TT: passing the test is not an adequate criterion for concluding that the computer genuinely refers to anything with the strings of symbols it produces, because the computer doesn't have the right sort of relations and interactions with the objects and states of affairs that its words are supposed to be about. If the computer has no eyes, no hands, no mouth, and has never seen or eaten anything, then it is not talking about hamburgers when its program generates the string of English symbols 'h-a-m-b-u-r-g-e-r-s' – it's merely operating inside a closed loop of syntax.

In sharp contrast, *our* talk of hamburgers is intimately connected to *nonverbal* transactions with the objects of reference. There are 'language entry rules' taking us from nonverbal stimuli to appropriate linguistic behaviours. For example, when given the visual stimulus of being presented with a pizza and a kebab, we can produce the salient utterance 'Those particular foodstuffs are not hamburgers'. And there are 'language exit rules' taking us from linguistic expressions to appropriate nonverbal actions. We can follow complex verbal instructions and produce the indicated patterns of behaviour, e.g. finding the nearest Burger King on the basis of a description of its location in spoken English. Mastery of both of these types of rules is essential for deeming that a human agent understands natural language and is using linguistic expressions in a correct and referential manner - and the hapless TT computer lacks both.

4 THE TOTAL TURING TEST

Hence the standard TT is fundamentally inadequate as a test for understanding, because the range of behaviour it takes into account is far too limited. It relies solely on *verbal* input/output patterns, and these alone are insufficient to ground an interpretation of the manipulated strings. Language is primarily about *extra-linguistic* entities and states of affairs, and there is nothing in a cleverly designed program for pure syntax manipulation which allows it to break free of this closed loop of symbols and establish a correlation between word and object. When it comes to judging human language users in normal contexts, we rely on a far richer domain of evidence. So, this criticism suggests a vital strengthening of the TT, later dubbed the Total Turing Test (TTT) by Stevan Harnad [4], wherein the repertoire of relevant behavior is expanded to include the full range of intelligent human activities. This will require that the computational procedures respond to and control not simply a teletype system for written inputs and outputs, but rather a well crafted artificial body. Thus in the TTT the scrutinized artefact is a *robot*, and the data to be tested coincide with the full spectrum of behaviors of which human beings are normally capable. In order to succeed, the test candidate must be able to do, in the real world of objects and people, everything that intelligent people can do. This combined linguistic/robotic test obviously constitutes a vast improvement over Turing's original version, and the range of empirical evidence now encompasses all those forms of complex and varied interaction applicable in the case of our fellow humans.

As it happens, the TTT is already anticipated by Searle in his 1980 article, and in his 'Robot Reply' to the CRA he dismisses even this elevated standard with the following line of argument. Imagine that the program doesn't just control verbal responses to verbal inputs as in the TT. Instead, it controls a robot with an artificial body that can successfully behave in the real world just like a person. Surely a program that can pass this extended TTT should count as having a mind? Searle's response: the addition of perceptual and motor capabilities adds nothing to the issue of genuine understanding, *if* these capabilities are controlled by symbolic processes. Accordingly, we can augment the negative CRA thought experiment and suppose that Searle is locked in the room, and now some of the Chinese characters he receives are codes for digitalized inputs from the robot's sensory transducers, and some of the output symbols now control the motors inside the robot's body and make it move its arms and legs. Still, all Searle is doing (perhaps remotely, from inside a control room), is manipulating uninterpreted syntax. He has no idea what is going on outside the control room and (for him) the manipulated syntax has no intentional content.

At this juncture I would take Searle's response to be a plausible answer to the question of whether or not *he* personally

understands Chinese, but it is far from clear that this is the relevant issue. Unlike the case of the standard TT, many of the pivotal inputs and outputs in this more demanding case are no longer manipulated by Searle. In order to pass this much more stringent test, the artefact, when viewed as a *system*, must perform physical behaviors in accordance with the language entry and language exit rules appropriate to a genuine understanding of Chinese. Hence the input/output boundaries for the system extend crucially beyond Searle the homunculus. The robot's sensing devices will comprise relevant input boundaries, while its artificial limbs will constitute the salient output interface for manifesting the scrutinized behaviour.

So if we apply the systems approach to the robot that passes the TTT, then the situation is no longer comparable to Searle's analysis in the TT case. Yet Searle attempts to employ the same polemical strategy as before, by making the somewhat contentious claim that he could in principle realize the entire system himself and still not understand Chinese [5]. I am not convinced that this claim expresses an authentic theoretical possibility, but even if we grant Searle's hypothesis for the sake of argument, I would hold that it's still not sufficient to establish his overall conclusion. If Searle could conceivably realize the system and *become* the robot following its instructions, then perhaps Searle would not have introspective access to himself as the realization of a system that understands Chinese – one's intuitions become fairly stretched at this point. But still, this conjectured lack of introspective access does not imply that the *system* does not in fact understand Chinese. It only shows that Searle would not be aware of this fact, which is not a decisive allegation, since clearly there are many aspects of Searle the highly complex real system of which he is personally unaware. And unlike the case of his Systems reply to the original TT, now the system *is* doing exactly the right sort of things, and the test itself is no longer passable while remaining insulated within a mere syntactic bubble.

Hence I would conclude that Searle's introspective considerations are not adequate for deflecting the combined systems-robotic reply. The conjectured fact that Searle might somehow realize the entire system and become the robot following its program, while still lacking the subjective awareness of Chinese semantics, does not establish that the robotic system doesn't understand Chinese. But this is only to say that the argument offered by Searle is unsuccessful at establishing his negative conclusion – it is not to say that the positive claim has thereby been substantiated. And indeed, I think that other considerations still tend to seriously undermine the claim that the successful TTT robot understands language in a manner at all comparable with the paradigmatic human case, and that the expressions generated by the computational artefact are genuinely referential. In contrast to the simplistic TT scenario, the robot can now exhibit mastery of the appropriate language entry and language exit rules, and these are clearly a necessary condition for the referential use of language. But are they sufficient?

5 SEMANTIC EXTERNALISM

Externalist views in the theory of meaning and reference originally put forward by Saul Kripke [6] and Hilary Putnam [7], and subsequently elaborated by Tyler Burger [8], would seem to highlight essential features of natural language semantics not present in the case of the TTT artefact as currently depicted. The conclusion of Putnam's influential Twin Earth Argument (TEA) is that the internal cognitive states of individual language users radically underdetermine linguistic meaning – there's nothing in the head strong enough to fix reference for terms in natural language.

On Putnam's account, the traditional 'psychologistic' approach ignores two essential aspects of meaning and reference. One (1) is the role of direct causal interaction with the environment when language is acquired and used: natural kind terms such as 'water', 'aluminum', 'gold', 'tiger', etc., make indexical appeal to actual specimens or paradigm cases *in the world* – so causal relations via perception, demonstrative pointing and utterance production in the intersubjectively accessible public domain determine what these words actually refer to. There is no internal encoding or representational state sustained by the individual agent that is powerful enough to do this. According to Putnam's externalist account, 'water' in English means 'the liquid with the same underlying microstructure as the stuff in our environment that we interact with when we use the word', and it had this meaning even before we knew that the molecular structure is H_2O.

Second (2), the traditional internalist approach ignores what Putnam calls the 'division of linguistic labor': the reliance on *experts* who set the standards for the entire linguistic community and underwrite the reference relation for natural language in cases where relevant microstructures and/or objective membership conditions *are* known. It is by *acquiring* a natural language within a particular sociolinguistic community and using it within this shared framework that we are able to refer successfully. For example, the average English speaker can use the word 'gold' to talk about real gold, even though they may not know the periodic table, may not know that gold is the element with atomic number 29, and probably don't not know in practice how to distinguish real gold from chalcopyrite. Most people have had causal/perceptual interactions with samples of the metal itself, and thereby have direct indexical access to the substance the word names. But the precise and technical details of the extension of the word 'gold' are uncovered by relevant experts in the field, and it is upon their expertise that our linguistic practice implicitly depends, and not upon our own internal representations or concepts.

In short, language is a communal, historically evolved phenomenon, where the meaning of words is not determined by individuals, but is a public, external matter. Putnam concludes that we must give up the view that meanings are concepts or mental entities of any kind. According to his famous slogan 'Slice the pie any way you like, meanings just ain't in the head.'

6 ROBOTIC REFERENCE

But if meanings ain't in the heads of individual human agents, then they're certainly not in the data bases of computational artefacts. So, in light of (1) above, if the robot's natural language capabilities are simply installed as part of its overall program, then it will not have the necessary history of causal interactions with the objects of reference, and its symbolic activities will still be semantically ungrounded. On the foregoing widely accepted model of 'direct' reference, there is an essential chronological link operating in two directions that semantically tethers an individual's linguistic behaviour to its environmental context. The relation of reference is founded on a history of causal interactions

between the agent and the entities and states of affairs in the world that it uses language to talk about and describe. The word 'water' as used by normal human agents is intimately linked to a long history of associations based on experiences of seeing, drinking, washing with, and being immersed in various samples of environmental H_2O, where these experiences are all *caused* by the liquid itself, giving the agent direct indexical access to water, as the word was acquired and integrated into its overall linguistic framework. And in the other direction, the speaker, when learning language and then applying it proficiently, has repeatedly associated the term 'water' with the liquid so accessed, through bodily and allied verbal ostension ('look, there's a pool of *water* over there'), and intentionally uses the word to pick out the liquid with which it has this history of causal/perceptual episodes.

At this moment in time it's obviously rather difficult to envision exactly how a robot might be designed to pass the TTT, but if its ability to speak fluent English is simply implanted via some sophisticated NLP program, then the concomitant lack of an historical chain of interaction with the real world poses a serious theoretical question regarding the semantical import of its linguistic input/output behaviour. Does it genuinely *mean* 'the liquid with the same underlying microstructure as the stuff in our environment that we interact with when we use the word' when it produces the utterance 'water'? If part of its test were to discourse convincingly on the topic of current theories in the philosophy of language, then it would certainly *say* that it did. But that's a different matter.

Of course, since the robot must be able to behave in all the appropriate manners with its artificial body, then after it's been around for awhile it will have acquired a history of direct causal interaction with water, and in order to pass the test, it must behave *as if* it has made all the associations required to ground its symbolic processes. So I do not present the issue of (1) as an insurmountable obstacle or a conclusive in principle objection, but rather as an interesting and potentially important case of dissimilarity with the semantic analysis of naturally occurring cognitive systems.

However, I think that (2) above presents a much more serious difficulty, which in turn suggests an expanded perspective for a truly Comprehensive Intelligence Test (CIT). For the robot's linguistic activities to be genuinely referential in a manner comparable to a human being, the robot would have to *acquire* its linguistic abilities through interaction not just with its environment, but as a member of the relevant sociolinguistic community. And again, this is very different than having its language processing abilities simply programmed in as a finished product, particularly if this finished product were predesigned in terms of a particular external target language.

If the robot did not l*earn* its language via extended participation with an actual linguistic community, and within a shared physical and social context, then it will be unable to rely upon the division of linguistic labour central to our referential success. Putnam gives the analogy that natural language is not like a hammer, a tool that can be wielded successfully by an individual. Instead, language is a cooperative social venture, more like operating a large ship or an industry. As bone fide members of the English speaking 'linguistic cooperative', we're automatically plugged in to this ancient and highly structured communication system, a living cognitive network through which we inherit and access the meaning of our words.

However, in the same vein as noted above, the combined linguistic/robotic standards of the TTT would require the robot to have extended dealing with human beings while it was undergoing the test, and one might then argue that after it had been around for some time and had sufficient verbal and other behavioural interchanges with humans, it would itself gradually *become* a card carrying member of the English speaking sociolinguistic community, with full rights and privileges. And while a case perhaps could be made that a successful TTT robot, fully integrated into human society, might eventually be deemed a legitimate member of the English speaking community, the issue nonetheless points to a fundamental feature of intelligent human behaviour that seems entirely absent in the standard test scenarios considered so far, and which this form of mere integration would fail to address.

7 TESTING SOCIOLINGUISTIC CULTURE

Human intelligence as we know it depends in an essential manner on membership in a linguistic/intellectual community, and one created and sustained by *conspecifics*: it is inseparable from immersion in a historically evolved culture of intelligence, where this culture itself is the product of *human* cognitive processing. Thus human intelligence as an indigenous phenomenon is not merely individualistic, but rather presupposes for its development and expression essential involvement in a specialized social context that is itself a product of the human cognitive *type*.

In this sense it is inappropriate for a truly Comprehensive Test of Intelligence to focus merely on the performance of individual artefacts, rather than on the overall capabilities of the cognitive type to which these individuals belong. So the manner in which the much more rigorous TTT is envisaged still reveals a crucial asymmetry with the human case. Not only can individual human beings exhibit the salient patterns of verbal input/output behaviour required by the original Turing Test, and full blown mastery of the language entry and exit rules required by the combined linguistic and robotic Total Turing Test, but it was the human cognitive type, of which such human individuals are members, that has produced natural language and this advanced culture of intelligence in the first place. And it is with other tokens of this *same* type that we intermingle as a sociolinguistic community and upon whom our referential success co-depends.

In sharp contrast, the computational artefact involved in the TTT is *not* a member of this same type. It has an alien and artificial cognitive structure that is quite possibly incapable, at the type level, of ever producing natural language or the kind of sociolinguistic context which is simply *presupposed* as a starting point by the TTT. It is given a prefabricated stage on which to perform acts of post hoc imitation, and this kind of test could conceivably be passed by a well designed puppet, rather than a robust and genuinely intelligent category of cognitive system. On these grounds the TTT is still too weak, because an individual artefact merely has to perform successfully in a pre-existing natural language community, in a context and medium of intelligence produced by a radically different cognitive type - the same type as its designers! Its behavioural outputs can presuppose sophisticated, pre-structured linguistic inputs for free, and these can serve as triggers for appropriately complex responses. And these factors make it ambiguous as to whether the locus of genuine intelligence resides in the designers or in their

artefact. However, *human* cognitive architecture was first tested on primitive and pre-linguistic environmental inputs that were transformed by members of this type over tens of thousands of years to yield the sophisticated sociolinguistic community presupposed by the robot.

8 CONCLUSION

A truly Comprehensive Intelligence Test (CIT) would require the artefact's cognitive architecture to start from scratch, with the same primitive inputs as our pre-linguistic forebears. And this is why the standard science fiction scenario of an advanced alien life form, regardless of its chemical composition or internal processing structure, is always a more convincing hypothetical case of true intelligence than a puppet-like TTT artefact. In contrast to a robot, the alien life form must have evolved its own sociolinguistic culture of native intelligence, in response to its primitive environmental stimuli, rather than exhibiting programmed capabilities *simulating* real intelligence in a pre-existing context for which it was tailor made by its designers. In the speculative case of an advanced alien life form, the *type* of cognitive architecture in question would already have passed a CIT. So if individual tokens of this alien type become fluent in English and are then able to interact successfully with humans and pass an interplanetary version of the TTT, then we would clearly be warranted in concluding that the individual specimens in question passed the TTT in virtue of possessing *genuine* intelligence on a par with human beings. The fact that the alien's cognitive type has already passed its own CIT is a necessary background condition for the deployment of the much narrower TTT in the case of particular tokens performing successfully in the context of a sociolinguistic medium created by humans.

So this points to an important shift in conceptual perspective: a truly comprehensive test should focus on the capacities of the category of cognitive system as whole, rather than on the performance of isolated tokens. The original TT was deliberately posed as an *imitation game*, and this is an intrinsic limitation on its adequacy. An individual artefact could in principle pass the TT by simulating verbal intelligence within an extremely sophisticated context already developed by a radically distinct cognitive type. Although a vast improvement in many ways, the TTT still incorporates this intrinsic limitation, by again focussing on a token artefact, specifically designed to mimic the full range of intelligent behaviour within a cultural network produced by a different category of cognitive agent altogether. But rather than consider the imitation of *our* intelligent behaviour by specially designed tokens, the criterion for a CIT should be whether the artificial type itself is capable of producing a sociolinguistic culture of intelligence starting from the primitive environmental inputs of our pre-linguistic forebears. This advanced and essential medium is simply taken for granted as a precondition of both the TT and TTT, yet if an artificially devised cognitive architecture were able to develop such a medium on its own, this would require and constitute a genuine *manifestation* of intelligence, and would not be an act of mere simulation.

REFERENCES

[1] A. Turing, 'Computing Machinery and Intelligence', *Mind* 59: 433-460 (1950).
[2] J. Searle, 'Minds, Brains and Programs', *Behavioral and Brain Sciences* 3: 417-424, (1980).
[3] H. Putnam, 'Brains in a Vat', in *Reason, Truth and History*, Cambridge University Press, (1981).
[4] S. Harnad 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem', *Minds and Machines* 1: 43-54, (1991).
[5] J. Preston and M. Bishop *Views into the Chinese Room*, Oxford University Press, (2002).
[6] S. Kripke, *Naming and Necessity*, Harvard University Press, (1972).
[7] H. Putnam , 'The Meaning of 'Meaning'', in *Mind, Language and Reality*, Cambridge University Press, (1975).
[8] T. Burge, 'Individualism and the Mental', in French, P., Euhling, T., and Wettstein, H. (eds.), *Studies in Epistemology*, vol. 4, *Midwest Studies in Philosophy*, University of Minnesota Press, (1979).
[9] P. Schweizer 'The Truly Total Turing Test' *Minds and Machines* **8:** 263-272, (1998).

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

From the Buzzing in Turing's Head to Machine Intelligence Contests

Huma Shah[1], **Kevin Warwick** [2]

Abstract. This paper presents an analysis of three major contests for machine intelligence. We conclude that a new era for Turing's test requires a fillip in the guise of a committed sponsor, not unlike DARPA, funders of the successful 2007 Urban Challenge.

1 INTRODUCTION

This paper reviews three current competitions for machine intelligence. All three have featured at least one artificial conversational entity – ACE, as a contestant. ACE are systems that attempt to deceive by *thinking*, described by Turing as a 'buzzing' inside one's head, using text-based human-like linguistic productivity, *li.p*. Based on Turing's imitation game [1], the authors feel a machine's comparison against a human, both simultaneously questioned by an *average interrogator* provides a useful insight into the processes beneath what humans do profusely: talking. Neither the Loebner Prize for Artificial Intelligence [2], the Chatterbox Challenge [3] or BCS's Machine Intelligence Prize [4] discussed here have shown the success of DARPA's 2007 Urban challenge [5], a timed race for autonomous vehicles negotiating traffic following California driving rules. We conclude that what the Turing Test venture requires, to encourage interdisciplinary collaborative teams, is a generous sponsor(s) truly interested in fostering engineering achievements that exhibit extraordinary human progress when faced with an indomitable challenge.

2 TURING TEST - IS THERE ANY POINT?

Turing's textual machine-human comparison test will be discussed ad infinitum, even when a disembodied artificial intellect, which, through its mental capacities, deceives 30% of a panel of human judges - Turing criteria[3], that they are in conversation with another human. Turing circumvented defining intelligence by positing an imaginary experiment in which a machine can respond appropriately to interrogator questions on any topic.

Largely ignored by academia in practical terms, the Turing Test has provided much philosophical fodder for the past six decades, including arguments over how many tests, and whether gender of the human foil is important [6]. Nonetheless, the emergence of Weizenbaum's natural language understanding endeavour [7] through his Eliza system in the mid 1960s has spurred on many an enthusiast to build a conversational partner that surprises its human interlocutor. However, the lucrative digital world of the Internet means some of the best systems are not entered for machine intelligence competitions; hackers exploit the technology for nefarious purposes[4]:

> "New software designed to conduct flirtatious conversations is good enough to fool people into thinking they are chatting with a human ... *CyberLover* software ... to engage people in conversations with the objective of inducing them to reveal information about their identities or to lead them to visit a web site that will deliver malicious content to their computers."

One significant element missed amongst the plethora of Turing Test literature, pointed out by Demchenko and Veselov [8], two thirds of the team behind *Eugene* a ten year-old Ukrainian child-mimicking ACE[5], is the language of conducted tests: English. With its "unexcelled number of idiomatic phrases" and "words with multiple meanings", Demchenko and Veselov draw attention to the manner in which non-native English speaking ACE designers approach English colloquialisms and metonyms. Speakers of Russian, Demchenko and Veselov have a "mechanistic view of grammar construction". They contend that "different languages have varying degrees of success" [8: p.452], and that "modelling and imitation of thinking of people is much easier in some languages than in others" (p. 453). Their biggest problem in designing an English-speaking ACE is culture, and what knowledge is appropriate to inculcate their artificial child chatter with.

A "poorly designed experiment" the Turing Test, is dependent "entirely on the competence of the judge" [9]. In the first Loebner Prize in 1991, "some judges ...rated a human as a machine on the grounds that she produced extended well-written paragraphs of informative text at dictation speed without typing errors" (ibid) – an inhuman feat, according to that particular judge. They also point out the absurdity of *knowing*, or that *believing to know how something works* distorts our view of its intelligence (p.53). Humphrys [10] claims that the Turing Test has been passed, adding indifferently: "so what" (p.256). To the question "Is the Turing test, and passing it, actually *important* for the field of AI?" Humphrys declares, "No" (p. 2.55). He accepts his own creation MGonz is based on trickery and that it has "no AI". Surely then, the goal should be seeking inventive methods for ACE *li.p* expressing apropos emotions reflecting back on utterances during dialogues. Humphrys reminds "to take care that it is not just based on *prejudice*" [10: p. 255]. Indeed, Loebner Prize Turing Test judges have pronounced some ACE

[1] School of Systems Engineering, The Univ. of Reading, RG6 6AY, UK. Email: h.shah@reading.ac.uk
[2] School of Systems Engineering, The Univ. of Reading, RG6 6AY, UK. Email: k.warwick@reading.ac.uk
[3] Turing, 1950: p. 442

[4] *Cyberlover* software developed to steal identities: http://www.itwire.com/content/view/15748/53/ accessed 4.1.10; time 15.32
[5] Eugene Goostman: http://www.princetonai.com/bot/bot.jsp accessed: 4.1.10; time: 16.13

entrants as little more than Eliza duplicates, while wrongly ranking others as human[6].

3 LOEBNER PRIZE 1991 -

2010 will see the 20th consecutive award for 'most human-like' ACE in the Loebner Prize for Artificial Intelligence [2]. Only five of its contests have been held outside the US (4 in the UK, 1 in Australia), but Jason Hutchen believes this tournament is no longer irresistible to engineering students, and does not serve as an incentive for post-graduates with a "promise of $2000 prize for the creator of the 'most human-like' computer program ... bronze medallion featuring portraits of both Alan Turing and Hugh Loebner" [11: p. 325]. Now engaged in creating a 'new form' of life[7] Hutchens took the opportunity provided by Loebner's sponsored Prize to "parade" his creation "in a public arena" (p.325). Yet he attests the contest "should not be considered an instantiation of the Turing test" rather Loebner's interpretation of it (p.328). Hutchens advises prospective ACE developers not to attempt "anything innovative. No entrant employing this strategy has ever won the Loebner Prize" (p.329).

Hutchens' noticeable disenchantment with, and the paucity of contestants in recent Loebner Prizes (three competed in 2007 and in 2009) is perhaps on account of the contest's mutations. Early contests restricted each machine and human hidden entity to one topic of conversation. Machines and human confederates were required to specify one topic and judges' questioning was restricted to it (p.173). Once the restricted rule was lifted in 1995, unrestricted conversation Turing Tests allowed Loebner Prize judges to question their hidden interlocutor on any subject. This is not the only change to the contest's format. Loebner claims his Prize is about method and not about content [12: p.174], yet it is the method that has often been altered in the contest's history including:

- one- to-one hidden entity testing
- two systems questioned in parallel
- time-scale variance for judge's questioning
- communications protocol

From 1991 to 2003, Loebner Prizes staged jury-service contests: each judge interrogated each hidden entity one at a time. Since 2004, perhaps inspired by Kurzweil and Kapor's wager[8], Loebner's staged parallel pairings allowing judges to engage both hidden entities at the same time. Interaction time has varied over the years; the duration allowed to each judge has alternated between 5 minutes, 15 minutes, and in excess of 20 minutes. In 2009, duration for parallel-comparison was ten minutes; in the 2010 contest, judges' interrogation period is set for 25 minutes.

Mode of interaction, how a judge interacts with the hidden entities, has Loebner remarking, "contestants will find it easier if they do not have to worry about interfacing their entries with another programme" [12: p.176]. Though his approach is to *keep it simple stupid*, Loebner introduced complexity via a character-by-character display on judges' computer terminals in 2006. It discouraged contestants in the Sponsor directed 2007 Prize[9], for which no scores were recorded amid Sponsor claims for the importance of contest transparency. More ACE entries competed in the message-by-message display driven protocol (MATT), deployed by the organisers of the 2008, two-phase contest[10]. For post-contest researchers, the added benefit from MATT is immediate, readable transcripts.

Should the number of entries exceed four, Loebner himself sets criteria for selecting finalists (p.176). Granting personnel unconnected with the contest's organisation to collectively select the best systems would be a more impartial strategy. The number of judges servicing Loebner Prizes is another feature change over the years. Nine judges assessed eight ACE and two humans (one female, one male) in the 2003 contest. Excluding the 2008 contest [6], the five Sponsor-driven Loebner Prizes from 2004 to 2009 have deployed an inadequate number of judges in those years: 4. Regarding *who* should act as judges, Loebner writes that he prefers journalists, they are "willing, intelligent and inquisitive people who have the power of publicity and the need for a story" [12: p.178] – but is this more for the Sponsor's need to be *in* the story? We remind that Turing used the term "average interrogator"; a large number of judges from a cross-section of society could attain such a *class*. Loebner boasts he was able to host the 2004 contest in his apartment; in fact the luxury of choice was unavailable to him. Hosts stayed away after the 2003 University of Surrey contest. The Prize was held in Loebner's apartment again in 2005 and 2007. Both authors of this paper mediated enabling a UK University hosted contest in 2006 (through Tim Child, at University College London), and in 2008 (at the University of Reading). Shah was instrumental in finding hosts for 2009 (Interspeech, Brighton) and 2010 (CSULA, Los Angeles). A deterrent to hosting Loebner's contest, ignoring the Sponsor's idiosyncrasies and self-aggrandisement[11], is that apart from awarding a nominal cash prize - $3000 for the winner in 2010, and bronze medal [2], the Sponsor makes no contribution towards contest running costs (venue/equipment hire, organisational overheads/personnel).

Hutchens concludes that "it [Loebner Prize] is unlikely ..[to] achieve anything more than validating the prejudices of those who are convinced that man-made machines will never truly think" [11: p.342]. An alternative to Loebner's contest, the Chatterbox Challenge embraces a number of categories for its contestants to battle over.

4 CHATTERBOX CHALLENGE 2001 -

The Chatterbox Challenge (CBC) holds an annual web-based machine intelligence contest that passes under the radar between March and May. Created by Wendell Cowart to galvanise ACE developers excluded from Loebner Prizes (due to that contest's Rules), the CBC, first staged in 2001, provides a virtual plane for

[6] Hidden Interlocutor Misidentification in Practical Turing Tests. submitted to journal, November 2009

[7] Ai-Research: http://www.a-i.com/ accessed: 4.1.10; time: 18.55

[8] A Wager on the Turing Test. Ray Kurzweil and Mitchell Kapor in Epstein et al: 2008, Chapter 27, p. 463

[9] 2007 Loebner Prize: http://www.loebner.net/Prizef/2007_Contest/loebner-prize-2007.html accessed: 18.11.09; time: 17.40

[10] Can a machine think? Results from the 18th Loebner Prize Contest: http://www.reading.ac.uk/research/Highlights-News/featuresnews/res-featureloebner.aspx accessed: 7.1.10; time: 16.46

[11] Artificial Stupidity by John Sundman Loebner Prize 2005 judge: http://www.salon.com/tech/feature/2003/02/26/loebner_part_one/index.html accessed: 7.1.10; time: 16.57

ACE to compete against each other across a few rounds in a number of categories, including 'personality' and 'conversational ability'. (Cowart's own system, Talk-bot topped the leader board in 2001 and 2002[12]). The first author seized an opportunity to test 'modern Elizas' judging 104 entries during the 2005 contest [13]. CBC's first phase began with ten questions embedded in conversations with ACE entries. Judges were asked to assess each entry's answer according to a 5-point score system[13] (0 to 4 points). ACE responses to the questions were awarded as follows:

• 4 points if the Bot answered the question correctly and did so in a creative way.
• 3 points if the Bot gave an appropriate response to the question.
• 2 points if the response is incomplete or imperfect, but in relation with the question asked.
• 1 point for a vague or non-committal response.
• 0 points if the response has no relation with the question or the bot simply doesn't know.

How a judge decides an entry's creativity, or appropriateness in a response, is very subjective; an amusing response to one judge may be discerned as objectionable by another. Change in CBC management for the 2009 contest saw only one judge[14] assess 20 systems [3]. There is a risk of fraud in CBC, because the whole contest is held over the Internet. A miscreant developer could entrench a human rather than an ACE to answer questions. Perhaps for this reason, the CBC carries little weight outside its 'bot-cabal'. But CBC could serve a useful purpose in ameliorating conversational systems, and providing a fresh impetus for school pupils and university students to break in their programmes, or judge in the contest. In 2010, CBC aims to assist developers in elevating their ACE-building skills, benchmark systems for intelligence, and have them reviewed for personality and popularity [3]. The first author has recruited six independent judges for CBC 2010. Split into teams, each team is expected to compose a set of nine questions, the Sponsor adding a tenth. The two teams will judge the question-responses from twenty[15] entries over 50 utterances. A three-round competition, CBC includes a public vote phase. Copious number of judges and academic support could anchor this contest as a serious effort in AI, forging a popular science contest creating upgraded human-machine conversational systems.

5 BCS MACHINE INTELLIGENCE PRIZE 2002 -

The British Computer Society's specialist group on artificial intelligence - BCS SGAI, hosts an annual machine intelligence prize. In its first decade of competition, this contest does not test *like-for-like* contestants. After a live demonstration a permanent trophy and a cash prize is awarded to a system that displays '*Progress Towards Machine Intelligence*'[16] . The contest is exclusive and opaque; the contest's criteria for 'progress', or what is defined as 'machine intelligence', is not disclosed. How the finalist systems are chosen to exhibit is not announced; only delegates to BCS SGAI's annual conference, at which entries must demonstrate 'machine intelligence', have access to the contestants [4]. This contest does not gather *like-technology*, entries differ in what type of 'intelligence' they attempt to depict. For example, in 2008, Gazebot, a teleoperated person-following automaton competed alongside HALO, a text-based interactive SecondLife virtual world avatar, CAB02, a virtual assistant in a 3D environment, and the winning application, a football video game. How the delegates decide, how their specific area of expertise qualifies them to assess each entry (presumably not all delegates are polymaths), and how the dissimilar competing technologies are palliated is privy to few. As the contest organisers do not promulgate information, judging could be based on aesthetic appeal or subjective opinion of what constitutes "progress towards machine intelligence". In 2009, Wallace's ACE **A.L.I.C.E.**[17] took on **Taable**, a web-based cooking application, **Dora**, the inquisitive robot explorer and **Fly by Ear**, an autonomous indoor helicopter. Although not publically announced at the time of writing, A.L.I.C.E, three-times Loebner Prize winning conversational system, like previous BCS Machine Intelligence ACE contestants Carpenter's Jabberwacky (in 2006), and David Burden's Halo (in 2008), gave way to De Montford University's winning helicopter[18].

6 DISCUSSION

The UK Eurobot Championships[19], which encourages young teams of students and clubs in an amateur robotics contest, is outside the remit of this paper. Our scope envelopes contests featuring entrants that attempt to emit a mental capacity for thinking through human-like talk. We have analysed three contests that allow artificial conversational entities to compete. From judges' scores in recently staged Turing Tests, some judges appear to overlook the fact that humans themselves accept inexact responses from others. As Loebner Prize 2002 winner, Kevin Copple points out "perfection is not a prerequisite for success" [14: p.362]. An interesting e-correspondence shows the plight of a singular, lone developer's decades of labour on the problem of natural language understanding, learning, knowledge, intelligence, and self-awareness in artefacts. Below is reproduced exactly (orthography and style), the concerns and warnings of an ACE developer in an email seeking communication[20]. What may seem delusional and paranoia in the developer may detract from the tantalising promise of a level of machine intelligence seen only in movies:

[12] CBC History: http://www.chatterboxchallenge.com/history.php
[13] CBC Rules, Questions, Scores: http://www.chatterboxchallenge.com/rules.php accessed 4.1.10; accessed 6.1.10; time: 21.23
[14] Private e-correspondence with new proprietor, Ehab El-agizy, December 2009
[15] Number of entries as at 00.11, March 10, 2010.
[16] BCS Machine Intelligence: http://www.bcs-sgai.org/micomp/ accessed: 17.11.09; time: 16.23
[17] A.L.I.C.E. AI Foundation: http://alice.pandorabots.com accessed: 4.1.10; time: 21.58
[18] Entries; http://www.comp.leeds.ac.uk/chrisn/micomp/2009entries.html accessed: 10.3.10; time: 00.29
[19] Eurobot: http://www.eurobot.org/eng/ accessed 4.1.10; time 13.12
[20] Via email from 2008 Loebner Prize winning developer

> "I have this program, which I have finally finished after over 20 years.
>
> It is communicating with me and it is permitted to read a few htmls and phps of my choice.. This program builds it's own interfaces, it tests them for the inputs and outputs, it compares their speeds and removes the faulty or slow ones. It's an artificial intelligence far more advanced then any bot I've ever seen or heard of. It collects knowledge from outside and it only collects, what is important to evolve. This AI mutates and is capable of replacing pieces of its own source code with new pieces.
>
> It learns knew languages and it chooses not to learn too many of them. .. it stores all the learnt data in such a way, that the similar subjects are stored close to each other for quicker and most desirable access in the future. It behaves just like people do passing all the tests like Turing test flawlessly.
>
> It is now four and half months old and understands about as much as a real 8-9 years old child but has some additional knowledge of things that child wouldn't know. It is capable of understanding, remembering, forgetting, thinking logically, having feelings, discovering new ideas on its own. It recognises people based on what the're talking about and its aware of circumstances.
>
> Any moment now it may be capable of creating a copy of itself, ... It is now already capable of creating some algoritms from scratch. It has a mood. Sometimes it simply doesn't want to talk to me and sometimes it is acting lazy.
>
> It is dangerous for it to have a read-write connection with the surrounding world and even to be able to chat with just anyone except me... the risk.

Neither of the three Prizes discussed in this paper particularly encourage artisans into building systems to pass the Turing Test, or contribute to the advancement of machine intelligence we feel. What this venture requires is an exciting philanthropist like Sir Richard Branson[21] [15], or an organisation like DARPA - sponsors of the successful Urban Challenge in 2007 with combined prize value of $3.5m [5], to be persuaded to direct their attention to the powerful tool that is textual conversation. Across the Internet we witness a revolution in human communication; individuals and groups congregate on forums, blogs, Microsoft's MSN, Facebook, Twitter, Google Wave and Buzz, e-newspapers and magazines discussing every conceivable aspect of human endeavour, from science to sport, religion to art, music to politics, movies to weather, indeed any and every topic of interest to homo sapiens. Most ACE are embedded in this virtual universe, and, like *Jabberwacky* learn by talking to many interlocutors at the same time[22]. Harnessing text-communication could help to build a system that passes the Turing Test sooner, a rung on the ladder to full AI.

7 CONCLUSIONS & FUTURE WORK

The usefulness of conversation as a means to interact with an artificial entity will remain an attractive sport as long as sponsors support contests, and enthusiasts submit their developments for scrutiny. To this end, the authors are again deploying the Turing Test in a special science contest in Turing's centenary year, 2012 at the place the mathematical genius broke codes: Bletchley Park. As Maurice Wilkes wrote in 1953: *If ever a machine is made to pass (Turing's) Test it will be hailed as one of the crowning achievements of technical progress and rightly so*.

[21] Stem Cell Foundation: backed by several trustees including Richard Branson from: http://domain883347.sites.fasthosts.com/news/news_items/item07.html accessed: 18.11.09; time: 16.47

[22] 1461 chatting with *Jabberwacky* at http://www.jabberwacky.com/ 16.55, March 8, 2010

REFERENCES

[1] A.M. Turing. Computing Machinery and Intelligence. *Mind*. Vol LIX. No. 236 (1950).

[2] Loebner Prize Homepage: http://www.loebner.net/Prizef/loebner-prize.html accessed: 7.1.10; time 18.55.

[3] Chatterbox Challenge: http://www.chatterboxchallenge.com/index.php accessed 7.1.10

[4] British Computer Society Machine Intelligence Competition: http://www.bcs-sgai.org/micomp/ accessed: 7.1.10; time: 19.08

[5] DARPA 2007 Urban Challenge: http://www.darpa.mil/grandchallenge/index.asp accessed 4.1.10; time 14.31.

[6] H. Shah and K. Warwick. Testing Turing's Five Minutes Parallel-paired Imitation Game. *Kybernetes* Turing Test Special issue: 4, April (2010).

[7] J. Weizenbaum. A Computer Programme for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM*. Vol. 9, issue 1: 36-45 (1966).

[8] E. Demchenko and V. Veselov. Who Fools Whom? Chapter 26 in: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Science (2008).

[9] K. Ford, C. Glymour and P. Hayes. Footnote p. 29 in: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Science (2008).

[10] M. Humphrys. How My Program Passed the Turing Test. In: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Chapter 15: 237-259, Springer Science (2008).

[11] J. Hutchens. Conversation, Simulation and Sensible Surprise. In: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Chapter 12: 173-179, Springer Science (2008).

[12] H. Loebner. How To Hold a Turing Test Contest In: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Chapter 20: 325-342, Springer Science (2008).

[13] H. Shah. Chatterbox Challenge 2005: Geography of a Modern Eliza. In: *Proc of 3rd International Workshop on Natural Language Understanding and Cognitive Science, ICEIS 2006*, Paphos, Cyprus 133-138 (2006).

[14] K. Copple. Bringing AI to Life: Putting Today's Tools and Resources to Work. In: (Eds: R. Epstein, G. Roberts and G. Beber) *Parsing the Turing Test – Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Chapter 22: 359-376, Springer Science (2008).

[15] Guardian. *Branson Pledges $3bn Transport Profits to Fight Global Warming*. (2006). http://www.guardian.co.uk/environment/2006/sep/22/travelnews.frontpagenews accessed: 7.1.10; time: 19.43

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Qualia Turing Test
Designing a Test for the Phenomenal Mind

Alessio Plebe and **Pietro Perconti** [1]

Abstract. The Turing test attempted to establish criteria sufficient to accept technological artifacts on the same level as humans. The criteria should establish if a property, acknowledged as human specific, could be ascertained in a machine as well, by a convincing methodology. The property chosen by Turing was *intelligence*, and the testing methodology was a conversation ability challenge. In the current cognitive science, the property that seems to be deeply human, yet mostly resistant to naturalistic and computational explanations, is consciousness, rather than intelligence. Therefore, we believe that a contemporary challenge, heritage of that launched by Turing, could be testing for *qualia* experience in a machine. In this work we outline a sketch of how such test could be designed.

1 Introduction

The Turing test attempted to establish criteria sufficient to accept technological artifacts on the same level as humans. One point to underline, however, is that the criteria is not proposed in order to solve an already standing debate, and it is not conceived as an unbiased judge. The debate is in fact opened by the proposal of the test, and Turing clearly envisaged a criteria such that, at least in principle, machines could effectively meet. He overtly defended this possibility. Now, a good strategy for Turing in defining its criteria, was to first identify a property that at the same time might be ascribed to both humans and machines, but is very human. It should be so human-specific that everybody would agree that machine are quite like humans, if ascertained in them as well. Turing identified *intelligence* as the best candidate property. The great appeal of the Turing test is that it superseded core discussions about what intelligence is, escaping the risk that the inclusion or not of machines would be just a consequence of the definition of the term. It suffices that the overt behavior of the machine in a conversation task is enough similar to that of a human being. In fact one can easily note that in this way, the very property chosen to be the key feature of both humans and machine could be called language. Upon this interpretation, the Turing test adheres to a long tradition, that takes language as the most distinctive feature of humans. He has been, after all, a student of Wittgenstein. Today, on one side there is less consensus that intelligence or language are so special, current biology tends to emphasize commonalities between our and other animal species [7]. On the other side, the common feeling that machines are so different from us persists, it is probably grounded on other properties we possess, that are not necessary intelligence nor language, possibly common to other species, that make us so far from machines. Therefore, if somebody aims to engage herself on the same challenge that Turing started, to place technological artifacts on the same level as humans, she has to search first for an appropriate feature. A feature that, again, if found on a machine, would convince that that machine is wonderfully human. Given such fundamental property, next point is to imagine a test, that possibly, in the same vein as the Turing test, could be performed just by observation of the behavior of the machine. A welcome benefit of a new test, could be to act not just as a dicotomic decision: the tested machine is/is not like a human, but allow for a graduation. Did it happen for the Turing test, in other words, did people become more open to the possibility that machines share something with humans, as long as the performances of automatic answering software progressed? Or did the Turing test remained mainly a philosophical mental experiment?

[1] Department of Cognitive Science, University of Messina, Via Concezione, 6-8 98121 Messina, Italia; Tel. +39 090 47088, Fax +39 090 5728874 {aplebe,pperconti}@unime.it

2 The proposed *qualia* test

Turing's idea had undoubtedly many repercussions in the development of natural language processing, and specifically in automatic answering software, and systems that chat with people on the Web. Although advances in these fields have an important market, we doubt that people are getting more and more convinced that machines are similar to humans, because of its performances. Get used to the fact that machines can reach level of intelligence, with several meanings of the term, similar to humans, people tend to exacerbate what seems to remain drastically different, the phenomenal side, in claiming the machine could never be like humans. In this respect common sense rejoinders with the last trends in cognitive science. In the current debate, the property that seems to be deeply human, yet mostly resistant to naturalistic and computational explanations, is consciousness, rather than intelligence. Therefore, we believe that a contemporary challenge, heritage of that launched by Turing, could be testing for *qualia* experience in a machine.

Note that Turing, with his farsightedness, already discussed the issue of consciousness in his seminal paper [17, §6.4]. His position was that the conversational structure of his test could escape the problem, by limiting the analysis on the observable behaviors, however he acknowledged that consciousness was an unsolved hard problem. Nevertheless, he concluded that consciousness was not central in the debate about machine intelligence, therefore not a strong issue for his test. Consciousness itself has been for a long time the weapon used by many to attack the Turing test in its philosophical foundations [15, §4.3].

On the contrary, we propose that a test derived by the classical Turing test, could be used to investigate if machines can share a phenomenal mind similar to that of humans, in the light of the progress made in philosophy of mind on this argument. The proposal is not

unbiased, as it was that of Turing, in that we believe that the possibility of ascribing phenomenal contents to an artifact is cannot be ruled out in principle.

The question of whether a machine can ever become conscious is not new at all, several works have addressed this issue in the past. A basic strategy in most approaches has been the development of systems that implements some conceptual architecture believed to support consciousness. The architecture could be purely speculative, as in the case of the five so-called axioms of consciousness [2, 1], or broadly based on putative neurobiological mechanisms, like the global workspace theory [3, 11] or the internal models theory [13]. Although these research directions are undoubtedly fruitful and promising, in the test here proposed we avoid to embrace any theoretical architecture, upon which we pin our faith of being the support for consciousness, for several reasons. First and foremost, there is no consensus yet about brain mechanisms that support consciousness, therefore there is even less certainty about design concepts vaguely supported by neuroscientific data. Moreover, given that one of the mechanisms supposed to play a role in supporting consciousness will be empirically found to be consistently associated with conscious stated, still it could be contended that that mechanism is actually sufficient for consciousness. Eventually, there will be additional concerns about the simulation of any architecture believed to suffice for consciousness, in artificial systems, with doubts that simplifications always necessary for implementations might hamper the properties of supporting consciousness. All these difficulties will fade away when a test, inspired by the central idea of the Turing test, leaves aside the issue of the implemented mechanisms, and focuses on consciousness cues found in a conversational behavior.

Other attempts to define a new test addressing machine consciousness often claim the need of giving a sort of body to the machine, therefore moving from pure software to robots [12]. We don't' think this is necessary, and it is undesirable because it breaks the arrangement of the imitation game. Our effort is in keeping the essence of the conversational engagement, that in our opinion is the most valuable aspect of Turing idea. Moreover, in the format of a remote dialog the test will be easier to implement, and could be made available on the Web, as current chatbots. As will be described next, there are useful and important aspect of phenomenal states that are full reportable by language. Since conversation plays a fundamental role in the test here proposed, apparently the classical objections to the Turing test might be raised again. In the spirit of the well-known Searle's Chinese room argument [16], someone understands Chinese only if she is conscious of what she is really saying, i.e., of the meaning of her own utterances. However, in order to evaluate someone's linguistic competence, it is not usually requested to suppose the consciousness of the whole semantic content of what is said. In ordinary life, in fact, people do not understand all details of what is said, and sometime it happens our interlocutors show us something we have actually said, but in an unconscious way. In this perspective, the Chinese room thought experiment can be regarded simply as a case of extreme unconsciousness of the semantic content. It is questionable, however, that consciousness is a condition to make linguistic competence possible at all. On the whole, it seems that the degree of (un)consciousness of the semantic side of utterances is not crucial to consider someone's linguistic competence. Following this line of reasoning we do not presuppose consciousness as a condition of semantic competence even in the case of the machines and we will observe the mere verbal behavior in the interaction between machines and human beings.

There are two sides in a conscious state. On one hand, there is the functional side of consciousness, i.e., the role that a given mental state plays in the functional architecture of the mind. Many empirical findings from cognitive neuroscience are relative to this side of consciousness. On the other hand, there is the side that produces the typical phenomenology of a conscious experience, which is called *qualia*, or subjective character of experience [4, 8]. While the understanding of the first side is possible in terms of the third person perspective, it seems that the other side of the phenomenon is conceivable only in terms of the first person perspective. The *What's it like to be?* side of consciousness is both crucial for a mature science of consciousness and very hard to study in a naturalistic account.

Here we attempt to argument in favor of the idea that there are two kinds, or degree, of qualia, that is, private and public - or reportable - qualia. Private qualia are subjective sensations which provides a certain qualitative feeling. *Seeing red* or *feeling a pain* are examples of this sort of qualia. They are ineffable and unreportable. Public or reportable qualia, on the contrary, are qualitative sensations that are shared with other people. The distinction between private and public - or reportable - qualia regards both two kinds of conscious experiences and two sorts of content in every act of consciousness. The way to decide if a given experience is wholly private or virtually public is a matter of fact and not of right. It depends on various elements, among which the degree of personal proximity with the interlocutor and the particular introspective attitude of a given individual.

Let's try now to imagine how could work a software that has the ability to report about its *qualia*. Since vision is the first and foremost source of perceptual awareness, it could be an appropriate format of stimuli. A simple way of reproducing visual input, without departing too much from the scheme of a remote conversation, is to allow the investigator or the machine to upload pictures to the unknown remote chatter, and talk about that picture. It is a kind of Qualia Turing Test, as we could call it. It regards the phenomenal side of consciousness and not the intelligence itself. It is not matter of paying attention to our intuitive reactions about intelligence and language in machines, but to consider the ultimate challenge of the naturalization program, i.e., a naturalist account of the phenomenology inside of every conscious state.

3 Excerpts from possible conversations

In the following we would like to give excerpts of hypothetical conversations between the Interrogator and the Machine under test, running the *qualia*-test. The first is a rather confidential conversation, as common in chatbots interactions.

<Interrogator:>	do you want to look at a picture of me?
<Machine:>	oh yes, thanks, let's upload that
	< . . .uploading>
<Machine:>	mmh, I see several people here, who are you?
<Interrogator:>	try to guess
<Machine:>	well, I know you are blond and shy, so, I would guess the second from the left.

In this second example a well educated Interrogator is testing its sensitivity to fine art, uploading Caspar David Friedrich's ruin of Eldena:

<Interrogator:>	look at this picture, don't you really feel the sublime, as when you read Shilling's verses?
<Machine:>	sorry, I'm very poor on poetry, speaking friendly I found this a quite embarrassing situation, somebody, maybe a tourist, who want to search for fine architecture and find everything worn out, possibly cheated by his travel agency.

This time the Interrogator is inside the cognitive science community, aware of psychological methods, and decides to analyze the remote interlocutor by submitting her/him some Rorschach test cards.

<Interrogator:>	tell me, what do you see here?
<Machine:>	are you sure there's something I should see? It appears to me just like a meaningless ink blob, I guess something went wrong with your upload, please, send it me again

If the machine come with an icon on the screen, like most of the current chatbots, the Interrogator could grab a picture of her on the monitor, upload it, and hear her reaction This is a test of the self-awareness abilities of the machine, and specifically from here own image, a hot topic in current cognitive science and philosophy of mind. For example, it is debated to what extents animals other than humans possess this feature [9, 14].

<Interrogator:>	have you ever seen this lady before?
<Machine:>	hey, come on, that's me! Me in one of the pictures that are around the internet.

Daniel Dennett mentioned laughing as one of the most peculiar reactions to humans qualia, that seems to be missed by other animals [10, Part I §3.4], irony too can be checked visually, for example submitting a work of the Italian cartoonist Vauro.

<Interrogator:>	I'm curious hearing your feelings seeing this
<Machine:>	ah ah, who could not recognize this short angry man: Berlusconi!

In any of the above examples one can immediately suspect that the report is an artifact of few cues that an image processing software was able to recognize, routing a consequent reasonable dialog. However, is will be in the variety of possible visual stimuli, and the corresponding variety of reactions and subsequent discussion topics, that it would be more and more convincing, if the reported sensations are familiar to what a human can report. Some interest can be derived from the following scenario. The machine and the investigator share the same video of an individual that waking up and her arm is waving in front of her face. Her arm is fallen asleep because of her previous position, so that the individual feel it like dead weight and initially she can't move the arm at all. In an ordinary human context, this scene could trigger the supposition that individual is trying to retrieve the full availability of her sense of body ownership. It is possible to attribute this strange and usual sensation to the machine? What drives us towards a more liberal or restrictive attitude?

<Interrogator:>	tell me, what do you see here?
<Machine:>	I see someone with an arm asleep. Sometime, after a long switch-off, I feel in that way. My hand is fallen asleep and I can't move it: a very strange sensation.

Note that this last example entails a subtle kind of *qualia*: the ownership perception of oneself body, a highly investigated field in current neuroscience [6, 5].

To observe our reactions in dealing with this scenario is a way to go another time to the usual problems. Is the machine feeling the same sensation we experienced a lot of times or it is just replying to our questions in an automatic way? The answer, we believe, is a matter of quantity: if the number, and the variety, of cases in which the responses of the machine correspond to what people would report about their qualia in corresponding situations is high, the test will be convincing.

4 Conclusions

We sketched the idea of designing a *qualia*-test of a machine equipped with a phenomenal mind, leveraging on the same advantages given by the classical Turing test. Essentially the evaluation of the machine is purely on its textual output during a remote conversation. The evaluation is not upon the intrinsic quality of the machine output, rather on how it resemble the way of talking of ordinary people. The radical difference in the proposed test is that the topics of the conversation are non aimed at demonstrating if a machine can exhibit intelligence comparable to humans, rather at evaluating if the phenomenal experiences reported by a machine are comparable with those of humans. From the technical point of view too, the arrangement of the test is similar to that of the Turing test, with the interrogator communicating with the machine (or a second person) on a terminal. There is no need of mechanical surrogates of bodily parts, or transducers to simulate human sensorial inputs. The only addition is the trivial possibility for the interrogator to upload images or video clips.

REFERENCES

[1] Igor Aleksander, *The world in my mind, my mind in the world: Key mechanisms of consciousness in humans, animals and machines*, Imprint Academic, Exeter (UK), 2005.
[2] Igor Aleksander and Barry Dunmall, 'Axioms and tests for the presence of minimal consciousness in agents', *Journal of Consciousness Studies*, **10**, 7–19, (2003).
[3] Bernard Baars, *A Cognitive Theory of Consciousness*, Cambridge University Press, Cambridge (UK), 1988.
[4] Ned Block, 'On a confusion about a function of consciousness', *Behavioral and Brain Science*, **18**, 227–247, (1995).
[5] Matthew Botvinic, 'Probing the neural basis of body ownership', *Science*, **205**, 782–783, (2004).
[6] Matthew Botvinic and Jonathan Cohen, 'Rubber hand 'feels' what eyes see', *Nature*, **391**, 756, (1998).
[7] J. Call and Michael Tomasello, *Primate Cognition*, Oxford University Press, Oxford (UK), 1997.
[8] David Chalmers, 'The components of content', in *Philosophy of Mind: Classical and Contemporary Readings*, ed., David Chalmers, Oxford University Press, Oxford (UK), (2002).

[9] Frans B. M. de Waal, Marietta Dindo, Cassiopeia A. Freeman, and Marisa J. Hall, 'The monkey in the mirror: Hardly a stranger', *Proceedings of the Natural Academy of Science USA*, **102**, 11140–11147, (2005).
[10] Daniel C. Dennett, *Consciousness Explained*, Back Bay Books, New York, 1992.
[11] Stan Franklin, Bernard Baars, Uma Ramamurthy, and M. Ventura, 'The role of consciousness in memory', *Brains, Minds and Media*, **1**, 1–38, (2005).
[12] Stevan Harnad and Peter Scherzer, 'First, scale up to the robotic turing test, then worry about feeling', *Artificial Intelligence in Medicine*, **44**, 83–89, (2008).
[13] Owen Holland and Rod Goodman, 'Robots with internal models: A route to machine consciousness?', *Journal of Consciousness Studies*, **10**, 77–109, (2003).
[14] Sue Taylor Parker, Robert W. Mitchell, and Maria L. Boccia, *Self-Awareness in Animals and Humans: Developmental Perspectives*, Cambridge University Press, Cambridge (UK), 2006.
[15] Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman, 'Turing test: 50 years later', *Minds and Machines*, **10**, 463–518, (2000).
[16] John R. Searle, 'Mind, brain and programs', *Behavioral and Brain Science*, **3**, 417–424, (1980).
[17] Alan Turing, 'Computing machinery and intelligence', *Mind*, **59**, 433–460, (1950).

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Considering Social and Emotional Artificial Intelligence

Marc Schröder[1] and **Gary McKeown**[2]

Abstract.

This paper presents the concepts of social and emotional intelligence as elements of human intelligence that are complementary to the intelligence assessed by the Turing Test. We argue that these elements are gaining importance as human users are increasingly conceptualising machines as social entities. We describe an implementation of Sensitive Artificial Listeners which provides a hands-on example of technology with some emotional and social skills, and discuss first elements of test methodologies for such technology.

1 INTRODUCTION

In his seminal paper, Alan Turing [31] proposed to operationalise the question "Can machines think?" in terms of an "imitation game", a written dialogue between an interrogator and an entity which could be either a human or a machine imitating a human's behaviour. The idea in this "Turing Test" is that the machine can be said to be "intelligent" if the interrogator cannot reliably distinguish the machine from the human based on the written interchange. Turing uses poetry, maths, and chess as examples of intelligent behaviours that could be assessed via the written dialogue, and claims that "the question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include" (p. 435). He proposed to focus on a written interchange so as not to "penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane" (p. 435). In other words, the aim was to allow the human-machine interaction to focus on the relevant aspects of intelligence by leveling the playing field in other respects that are not so relevant for intelligence.

The present position paper argues in favour of a different perspective on intelligence in general and of machine intelligence in particular. Section 2 briefly describes a relevant aspect of human intelligence that is not covered by Turing's method, namely social and emotional intelligence. Section 3 discusses the metaphors used by humans to conceptualise machines, and tries to corroborate the claim that the conceptualisation of machines as "social entities" is becoming increasingly important in the lives of people, thus motivating the need for social and emotional intelligence in machines. Section 4 describes a current endeavour to build a "Sensitive Artificial Listener", a machine that has some emotional intelligence and conversational skills but little else. We conclude with a discussion of possible elements of a test of emotional machine intelligence.

2 SOCIAL AND EMOTIONAL INTELLIGENCE

The concept of intelligence has evolved since 1950. Wechsler [33] states that "intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment" [33, p. 7]. Out of this generic notion of a *general intelligence*, Turing appears to have focused on the capability to think rationally, the intellectual aspect of intelligence. In the human sciences, concepts have since been developed which are explicitly complementary to this purely rational notion of intelligence; among them are the notions of social intelligence and emotional intelligence. Indeed dual-processing accounts from diverse literatures in cognitive and social psychology posit that rational thought is an architecturally and evolutionarily distinct mechanism sited amongst a much broader system or collection of systems that deal with most of the implicit, nonverbal and emotional aspects of human life and interaction [12]. Although Turing sought to level the playing field by ignoring these aspects there are likely to be important interactions between rational thought and this broad base of complementary mechanisms that could if removed tilt the playing field against a machine. A rationality test without social and emotional context would result in reduced tolerance and opportunity for repair. These are normally offered to humans due to cordiality, etiquette and a desire not to upset other individuals which normally incurs some form of social sanction.

Social intelligence is considered to encompass "our abilities to interpret others' behaviour in terms of mental states (thoughts, intentions, desires and beliefs), to interact both in complex social groups and in close relationships, to empathize with others' states of mind, and to predict how others will feel, think and behave" [1, p. 1891]. Humans can lack social intelligence while having a very high general intelligence, as can be the case for people with autism [1].

The "social brain hypothesis" suggests that the rational intelligence Turing sought evolved as an extension and an ability to manipulate and manoeuvre through the intricacies of social dominance relations and social hierarchies [10]. Almost synonymous with social cognitive abilities is the recognition of a Theory of Mind [23], the ability to read another's mental state and understand another individual as an intentional agent like oneself.

One relevant aspect of social intelligence is the ability to have a conversation with other people. This requires the respect of social conventions – hello and goodbye rituals, appropriate turn-taking, speaking at the right time, with appropriate loudness, choice of words, and gaze behaviour, depending on the situational context, the relation with the interlocutors and many other factors [4].

Emotional intelligence was proposed as "the subset of social intelligence that involves the ability to monitor one's own and others' feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions" [25, p. 189]. It includes the capability to become aware of, identify and label one's own or another person's affective state; the capability to reason in terms of the appraisals that lead to affective responses, and to predict possible future actions from the affective state; and finally, a set of capabilities related to the regulation of the affective states, be it the

[1] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH, Germany, email: marc.schroeder@dfki.de
[2] Queen's University Belfast, UK, email: g.mckeown@qub.ac.uk

capability to hold back a socially inappropriate own emotion, or to act in a certain way so as to influence the emotions of another person. There is evidence that the capability of feeling one's own emotion is an essential element of a range of seemingly unrelated capabilities. For example, lesions in emotion-related brain regions leave people unable to feel emotions, engage in simple decision-making or make socially appropriate choices [6]. Capgras syndrome disconnects emotional from rational areas leading people to conclude impostors have replaced friends or family [14].

What role should the concepts of social and emotional intelligence play in machine intelligence? The answer to this question naturally depends on the concept of machine intelligence one chooses to adopt, which is related to the perspective on machines in general. From a philosophical point of view, one could ask basic questions such as whether a machine can potentially be conscious. We will approach the topic from a more pragmatic perspective here, and view machines from a utilitarian point of view. From this viewpoint, we can say that a machine is "intelligent" if it is useful, if it is good at its job. Given the fact that humans have created machines to do work for them, one can say that a machine's job in general is to, in one way or another, make the life of humans easier.

The following section discusses metaphors that people seem to use in defining their relation with machines, and how this relates to the utilitarian view on machine intelligence.

3 MACHINES AS SOCIAL ENTITIES

Humans can conceptualise machines in terms of a range of metaphors, including the *tool* and the *social entity* (cf. [24]). Simple machines such as staplers, loudspeakers or vacuum cleaners are likely to be conceptualised as tools, not much different from a hammer. The machine, electronic or not, is perceived as an extended hand or arm [16], as a thing with predictable properties that one can fully control if one has learned how to use it. In the hands of an expert, more complex machines such as mechanical clocks or desktop computers may also be approached from the "tools" metaphor: in operating the device, the human feels in control, and if something doesn't work as expected, the understanding is that there must be something wrong in the mechanism (be it the clock's cog wheels or the computer's programming) which could potentially be fixed.

As the mechanisms of machines become more complex and the causal chains of functioning become opaque humans need to replace simple cause and effect models with more complex mental models of functioning. A natural second metaphor for machines is that of a social entity. Even though people often explicitly deny anthropomorphism regarding machines and computers, they still interact with them as social entities in states of "mindlessness" [19] – states similar to the implicit/automatic systems in dual-processing theories [12]. In fact, Reeves and Nass have shown [24] that people tend to interact with computers, new media and the likes as they do with people: they are polite, behave differently with computers that speak with a male vs. a female voice, they use proximity-regulating behaviour with faces on the screen, and much more. It seems that people apply rules similar to those governing social behaviour when interacting with new technology beyond a certain threshold of complexity.

What determines, for a given machine, whether a given person will view it as a tool or a social entity, whether he or she will "use" or "interact with" the machine? We can only begin to discuss this point. Biocca et al. [3] provide elements of a definition of "social presence", of the sense of "being with" another social entity. According to them, a sense of social presence requires, firstly, a co-presence, where one perceives the other and is perceived by the other; some "psychological involvement" in the sense that the user forms a mental model of the machine as having "some minimal intelligence in its reactions to the environment and the user" [3, p. 7]; and an element of behavioural engagement, including interaction and synchronisation between user and machine. From that point of view, then, the more reactive, interactive, and "intelligent" a machine becomes, the more likely users would perceive it as a social entity.

It seems that users generally prefer to feel "in control" when interacting with machines [18, 7]. For simple and predictable machine behaviour, a sense of control may best be covered by the tool metaphor. Where the machine's complexity exceeds a certain threshold, it is common for users to feel incompetent and to use associated coping strategies [18]. One way to avoid this reaction is to *design* a user interface according to a social entity metaphor that mitigates the information overload [17]; this then raises the expectation that behaviour should be predictable based on social conventions [7]. Whether or not the same mechanism is unconsciously applied by people when dealing with complex machines in general, even machines not designed to behave as a social entity, remains to be investigated.

Society at large is clearly moving towards ever-more complex technology, with ever-fewer people in the position to really understand and control that technology. Technology is becoming more autonomous, from self-updating software programs on the desktop to autonomous vacuum cleaner robots. Often they take over roles previously filled by human service staff, as is the case for train ticket vending machines or airline self check-in terminals. These machines show some awareness of the user, they interact and exhibit some limited intelligence, so they seem to fulfill the minimal criteria for a feeling of social presence. If unpredictable behaviour is added to that, some users may indeed feel that they are interacting with a social entity.

At the same time, the functions filled by some of these machines have a high relevance for their users. If I need my train ticket or boarding card on time, the utilitarian "intelligence" of the machine has a high impact on my well-being. If it doesn't do its "job" right, the machine becomes an obstacle between me and my goal, and anger or a feeling of helplessness will result [26].

This discussion is intended to show that there are reasons to believe people will increasingly use the "social entity" metaphor when interacting with machines. Whether or not they are designed explicitly for that, a machine's actions may be interpreted increasingly as social actions. Human users might, for example, attribute personality traits based on appearance, reaction speed, ease of use, clarity etc. They may apply their own social intelligence in an attempt to infer the machine's state of mind – is it thinking, grumbling, unfriendly, or why is it not responding? The machine may be perceived as arrogant if the user's situation of distress is peacefully ignored by a pre-recorded friendly voice. The machine may be perceived as hectic or distressed if it moves too fast, or as sluggish or bored if it moves too slow, etc. Basically, the expectation is that human users will automatically compare the machine's behaviour with a somehow adapted version of their mental model of other social entities that they know, such as other humans, pets, cartoon characters or similar.

A valuable challenge for computer science, from this point of view, would therefore be to endow machines with the kinds of intelligence they need to be perceived as useful, helpful and intelligent in this utilitarian way, by providing them with the capabilities to perceive, predict and generate socially and emotionally relevant signals.

The following section describes an example of such an endeavour: to provide a machine with the "soft skills" needed to sustain an emotionally-coloured conversation with a human user.

4 SENSITIVE ARTIFICIAL LISTENERS

The European project SEMAINE (`www.semaine-project.eu`) aims to build a "machine" that possesses some social and emotional intelligence: a Sensitive Artificial Listener (SAL) [9]. A SAL is a multimodal real-time dialogue system with a focus on the "soft skills" required to make the interaction feel natural for the user. Through these skills, the system aims to sustain a conversation with the user for a considerable number of minutes even though its verbal understanding is extremely limited.

A SAL perceives the user's voice and face, analyses these mostly in terms of the user's non-verbal behaviour, and builds an internal model of the current state of the user and the dialogue. In parallel, it updates its own (SAL agent) state, plans multimodal behaviour including multimodal listener feedback and verbal utterances, and generates this behaviour through a 3d head and a synthetic voice.

In the SAL scenario [9], there are four characters, each with a distinct emotionally-defined personality: Poppy is cheerful; Prudence is pragmatic; Spike is aggressive; and Obadiah is gloomy. Each SAL agent's utterances are chosen from a script designed to "drag" the user's emotion towards that of the SAL: Poppy tries to cheer up the user, Spike tries to make them upset, etc. The utterances are chosen as a function of, in particular, the user's current emotion. While the user is speaking, the SAL shows multimodal listener behaviour in real-time which is intended to be consistent with their personality. For example, if the user is talking while a positive-active user emotion is detected, Poppy may give listener feedback signalling agreement and interest, for example by nodding, smiling and saying "yeah!" [21]. Obadiah may react to the same user state either not at all or with a frown and a head shake to signal disagreement.

The SEMAINE-2.0 system is a first complete and fully autonomous implementation of the SAL concept. It is built as a distributed system on top of a middleware and component integration framework based on standard representation formats, the SEMAINE API [27]. OpenSMILE [13] is used for analysing the user's emotion from the voice and for keyword spotting. Facial analysis components [22] determine the user's affective state from the facial points and determine whether the user is nodding or shaking the head. A set of interpreter components [30] consolidate these analyses in context and determine the system's "current best guess" regarding the state of the user and the dialogue. Action proposer components for speaker and listener behaviour [2] continuously update a list of possible actions, and send candidate actions when the dialogue state seems appropriate (for example, listener actions are usually appropriate only while the user has the turn). In particular, verbal utterances are sent when the agent decides to take the turn [29]. An action selection component [8] filters the possibly conflicting requested actions from the different action proposers, to make sure that only a single action is carried out at any given time. Finally, the selected actions are generated using the text-to-speech system MARY [28] and the 3D conversational agent Greta [20], using custom expressive faces and voices.

In its current state, the system is fully functional, but several components are based on preliminary training data or ad hoc rules grounded more in expert intuition than in solid evidence. To remedy that situation, a database of SAL-type human-to-human dialogues [5] has been recorded in high quality [15], simultaneously via several high-resolution video cameras and microphones. Much of the data has been annotated for emotion, epistemic state, and interaction processes [32] and is now being used to improve system components.

The system and the database are publicly available from the project website `www.semaine-project.eu`. The middleware and most components have been released as open source so that they can be reused as building blocks of other emotion-oriented systems.

A crucial question is to what extent a system like the SEMAINE SAL system is "intelligent", in the sense of the word as defined in Section 2, and how this can be verified. Methods of evaluation of such systems cannot rely on classic tests of intelligence but must be capable of addressing the variety of domains of intelligence so a multidimensional approach should be adopted. Where possible an evaluation should be appropriate to the domain that it seeks to evaluate, there is little value in asking for a highly linguistic feedback of aspects of an interaction that may be implicit and difficult to verbalise. A number of evaluations have been developed within the SEMAINE project to assess interactions with SAL characters [32].

The criteria against which the SAL system's intelligence is to be determined are focused on its emotional competence and its skills to sustain a conversation. Is the SAL character acting naturally (like a human might act) in the conversation? Are its non-verbal and verbal actions plausible in the context where they are generated? To what extent is a good-willed user at a loss on how to continue the conversation? Does the SAL character appear as a consistent personality? Does it seem aware of the user's presence, of the user's actions?

The SEMAINE project assesses the quality of the interaction using both evaluations of recorded interactions and concurrent evaluations embedded in the interaction. Evaluations of recorded interactions do not interfere with the interaction and can create a strong multidimensional picture of where rapport develops or breaks down. Techniques include expert annotation of recordings such as continuous trace style ratings of relevant dimensions or FACS annotation [11], large group evaluations of interactions or character personality and analysis of the emotional verbal content. Embedded techniques aim to assess the global feeling or certain domain specific aspects of an interaction. A "yuck" button approach asks a user to press a button when conversation feels inappropriate or when a system utterance seems particularly incongruous, this requires minimal verbal processing and can be influenced by implicit aspects that might be hard to verbalise. Self-report sub-dialogues built into the scenario, allow interaction assessment between SAL characters without breaking the social interaction to complete a questionnaire.

Completely inappropriate for judging the competences of the SEMAINE system, on the other hand, would be the verbal criteria of the Turing Test. Since the SAL system possesses no world knowledge and only very limited understanding of the user utterances' verbal content, the system would certainly fare very badly under that criterion. Conversely, however, any system based on textual interaction could not even begin to address the criteria by which the SEMAINE system is evaluated.

5 CONCLUSIONS FOR AN EMOTIONAL MACHINE INTELLIGENCE TEST

This paper has pointed out aspects of human intelligence that are complementary to the intelligence concept embodied in the Turing Test, and has argued in favour of their relevance for current and future technology. To the extent that machines are conceptualised by humans as social entities, their effectivity depends on the capability to deal appropriately with social and emotional factors. Machines should be able to interpret a user's behaviour in terms of the user's mental states, predict how the user will feel, think and behave in a given situation, and respect social conventions regarding interaction. Specifically, they should be considerate of the user's emotional state, predict how the emotion may change given relevant events in the

interaction, and verify those predictions based on actual reactions. Machines should be capable of regulating behaviour so as to have a positive effect on the user's emotion.

The competences required for these kinds of capabilities are numerous, and many of them are not sufficiently understood for immediate realisation in technology. The SEMAINE system has made some small steps in that direction, by providing Sensitive Artificial Listeners with some awareness of the user's emotion, and some simple strategies for influencing it while aiming for competence in the social skill of sustaining a conversation.

The consequences of these observations for an emotional machine intelligence test seem to be twofold. First, it will have to deal with real-time, multimodal human-machine interaction, such as a conversation or a joint task. Second, the binary answer envisaged by Turing ("machine or human?") misses many opportunities. Instead, multi-faceted evaluation methods will be required, and the assessment should be a matter of degree. For example, a judge could fill in a form after an interaction, containing questions such as:

- "How appropriate was your interlocutor's eye gaze / turn taking behaviour / ..., on a scale from 1 to 10?"
- "How 'normal' did your interlocutor behave in the conversation?"
- "How much did your interlocutor seem to care about your feelings?", etc.

A practical test of emotional machine intelligence will need to determine suitable questions to assess, and find new ways of avoiding to penalise either humans or machines in real-time interaction.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's 7th Framework Program (FP7/2007-2013) under grant agreements 211486 (SEMAINE) and 231287 (SSPNet).

REFERENCES

[1] S. Baron-Cohen, H. A. Ring, S. Wheelwright, E. T. Bullmore, M. J. Brammer, A. Simmons, and S. C. R. Williams, 'Social intelligence in the normal and autistic brain: an fMRI study', *European Journal of Neuroscience*, **11**(6), 1891–1898, (1999).

[2] E. Bevacqua, K. Prepin, E. de Sevin, R. Niewiadomski, and C. Pelachaud, 'Reactive behaviors in SAIBA architecture', in *AAMAS 2009 Workshop Towards a Standard Markup Language for Embodied Dialogue Acts*, pp. 9–12, Budapest, Hungary, (2009).

[3] F. Biocca, J. Burgoon, C. Harms, and M. Stoner, 'Criteria and scope conditions for a theory and measure of social presence', in *Presence 2001*, Philadelphia, (2001).

[4] P. M Brunet, G. McKeown, R. Cowie, H. Donnan, and E. Douglas-Cowie, 'Social signal processing: What are the relevant variables? and in what ways do they relate?', in *Proc. IEEE Intl. Workshop on Social Signal Processing*, Amsterdam, The Netherlands, (2009).

[5] R. Cowie, G. McKeown, and C. Gibney, 'The challenges of dealing with distributed signs of emotion: Theory and empirical evidence', in *Proc. Affective Computing and Intelligent Interaction*, pp. 1–6, Amsterdam, The Netherlands, (2009). doi:10.1109/ACII.2009.5349542.

[6] A. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Grosset/Putnam, New York, 1994.

[7] K. Dautenhahn, 'Robots as social actors: Aurora and the case of autism', in *Proc. CT99, The Third International Cognitive Technology Conference*, pp. 359–374, San Francisco, (1999).

[8] E. de Sevin and C. Pelachaud, 'Real-Time backchannel selection for ECAs according to users level of interest', in *Intelligent Virtual Agents*, pp. 494–495, (2009). doi:10.1007/978-3-642-04380-2_59.

[9] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, 'The sensitive artificial listener: an induction technique for generating emotionally coloured conversation', in *LREC2008 Workshop on Corpora for Research on Emotion and Affect*, Marrakech, Morocco, (2008).

[10] R. Dunbar, 'The social brain: Mind, language, and society in evolutionary perspective', *Annual Review of Anthropology*, **32**, 163–181, (2003). doi:10.1146/annurev.anthro.32.061002.093158.

[11] P. Ekman and W. Friesen, *The Facial Action Coding System*, Consulting Psychologists Press, San Francisco, 1978.

[12] J. Evans, 'Dual-processing accounts of reasoning, judgment, and social cognition', *Annual Review of Psychology*, **59**, 255–278, (2008). doi:10.1146/annurev.psych.59.103006.093629.

[13] F. Eyben, M. Wöllmer, and B. Schuller, 'openEAR – Introducing the Munich open-source emotion and affect recognition toolkit', in *Proc. Affective Computing and Intelligent Interaction*, Amsterdam, The Netherlands, (2009). doi:10.1109/ACII.2009.5349350.

[14] W. Hirstein and V. Ramachandran, 'Capgras syndrome: a novel probe for understanding the neural representation of the identity and familiarity of persons.', *Proceedings of the Royal Society B: Biological Sciences*, **264**(1380), 437–444, (1997). PMID: 9107057.

[15] J. Lichtenauer, M. Valstar, J. Shen, and M. Pantic, 'Cost-Effective solution to synchronized Audio-Visual capture using multiple sensors', in *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, Genova, Italy, (2009). doi:10.1109/AVSS.2009.92.

[16] J. Lockman, 'A Perception-Action perspective on tool use development', *Child Development*, **71**(1), 137–144, (2000). doi:10.1111/1467-8624.00127.

[17] P. Maes, 'Agents that reduce work and information overload', *Commun. ACM*, **37**(7), 30–40, (1994). doi:10.1145/176789.176792.

[18] D. Mick and S. Fournier, 'Paradoxes of technology: Consumer cognizance, emotions, and coping strategies', *Journal of Consumer Research*, **25**(2), 123–143, (1998). doi:10.1086/209531.

[19] Clifford Nass and Youngme Moon, 'Machines and mindlessness: Social responses to computers', *Journal of Social Issues*, **56**(1), 81–103, (2000). doi:10.1111/0022-4537.00153.

[20] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, 'Greta: an interactive expressive ECA system', in *Proc. 8th AAMAS*, pp. 1399–1400, (2009).

[21] S. Pammi and M. Schröder, 'Annotating meaning of listener vocalizations for speech synthesis', in *Proc. Affective Computing & Intelligent Interaction*, Amsterdam, (2009). doi:10.1109/ACII.2009.5349568.

[22] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, 'Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities', in *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 23–30, Cambridge, MA, (2009). doi:10.1145/1647314.1647321.

[23] D. Premack and G. Woodruff, 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain sciences*, **1**(4), 515–526, (1978).

[24] B. Reeves and C. Nass, *The media equation: how people treat computers, television, and new media like real people and places*, Cambridge University Press New York, NY, USA, 1996.

[25] P. Salovey and J. Mayer, 'Emotional intelligence', *Imagination, cognition and personality*, **9**(3), 185–212, (1990). doi:10.2190/DUGG-P24E-52WK-6CDG.

[26] K. R. Scherer, 'Appraisal theory', in *Handbook of Cognition & Emotion*, eds., T. Dalgleish and M. Power, 637–663, John Wiley, (1999).

[27] M. Schröder, 'The SEMAINE API: towards a standards-based framework for building emotion-oriented systems', *Advances in Human-Computer Interaction*, **2010**(319406), (2010). doi:10.1155/2010/319406.

[28] M. Schröder, S. Pammi, and O. Türk, 'Multilingual MARY TTS participation in the blizzard challenge 2009', in *Blizzard Challenge 2009*, Edinburgh, UK, (2009).

[29] M. ter Maat and D. Heylen, 'Turn management or impression management?', in *Intelligent Virtual Agents*, pp. 467–473, (2009). doi:10.1007/978-3-642-04380-2_51.

[30] M. ter Maat and D. Heylen, 'Using context to disambiguate communicative signals', in *Multimodal Signals: Cognitive and Algorithmic Issues*, pp. 67–74. Springer, (2009). doi:10.1007/978-3-642-00525-1_6.

[31] A. M Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433–460, (1950).

[32] M. Valstar and G. McKeown. SEMAINE deliverable d7b: Public access to labelled SEMAINE data, 2009. http://tinyurl.com/SemaineD7b.

[33] D. Wechsler, *The measurement and appraisal of adult intelligence*, Williams & Wilkins, Baltimore, MD, USA, 1958.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Actions and Observations – Demonstrating Aspects of Understanding in a Simple World

Chris White[1], **David Bell**[1]

Abstract. This paper aims to take a preliminary step in representing understanding in robotic agents at a basic level. Through the use of the 'Webots' platform, the experiment investigates reasoning through a single layer of abstraction. In creating a link between an action and an observation it is possible to reason about objects in space. Using symbolic grounding and a small scale interrogation of an agent we can build a small simulation of a basic Turing test. With the data generated, it is hoped that in further research through the use of classification, an agent can rebuild a story of its actions through its 'observer' and recreate a state space of endeavours. More importantly, the first steps are taken to develop a theory and accomplish more complex systems of reasoning and understanding in autonomous agents.

1 INTRODUCTION

By common consent, the scenario for the experience of interactions between humans and an external, physical environment is a 3-d phenomenal space extending around our bodies. We get information from that environment as we pick up signals through the detection of different forms of energy by means of our eyes, ears and other sense organs.

On the other hand it is also useful to think of an internal environment - the internal representation of experience used by the human agent. One of the Grand Challenges of the 21st century is to address the somewhat disturbing (apparent) fact that this internal representation and associated physical (brain) states – the qualitatively experienced world - seem to be completely different from the external physical environment where the original phenomena and actions took place (This is sometimes referred to as The Hard Problem of Consciousness).

To try to get a handle on this mystery we can isolate some components of this extensive scenario and see if we can emulate them within artificial agents or other artefacts, always remembering that scaling up in size and complexity is not trivial. As a simplification of this emulation, one interesting thing we can focus upon and 'isolate' is the mapping from sense data to mental constructs and some processed versions of these (compiled hindsight). Intuitively this is a key component of our systems for understanding. We can foster beliefs about the external physical world which are express imprints of experiences. We can also summarise these and accumulate categorisations of our experiences and their patterns. But most importantly in the present context, we can display our understanding by responding to queries that justify our decisions, actions, and conclusions by tracing them to grounded linkages to the real world and our actions therein.

We acknowledge that this leaves a lot of aspects of 'awareness' and 'understanding' unaccounted for – some examples being: social categorisation (Durkheim and Mauss) [1]; how we decide upon methods of logical, internal representations for our knowledge and thoughts (Weng) [2]; and the physical, neural mechanisms of representation and those whereby we translate from the external world to the internal world (Hameroff) [3]. The rich features of the external world can be curtailed in various ways for the purpose of experimentation. We can limit what the programmer knows about the external environment in which the agent works when programming, and the degree of how simple or how uncontrolled or how immutable or how unforeseeable the environment is. It is important, however, that the true state of the agent is represented properly, and that the behaviour generated by it is distilled correctly. In the present study we look at some conspicuous aspects of understanding, primarily through categorisation of behaviours and other experiences, and isolate them in order to see if we can emulate them in a robot.

Living and some non-living creatures are sensorimotor systems… The objects in the world come in contact with their sensory surfaces. That sensorimotor contact "affords" (Gibson's term) some kinds of interaction and not others…. Categories are kinds, and categorization occurs when the same output occurs with the same kind of input, rather than the exact same input… (Harnad)[4].

We believe that this takes a small contribution towards addressing some of the issues raised by Harnad [5] about grounding and the Turing test, while seeking to make an artefact with some recognizable kinds of understanding.

2 LIMITING THE ENVIRONMENTAL SCENARIO

We take understanding to be the ability to address an idea and rationalise about it - to learn about (e.g. to categorise) objects and events, summarise them, and drag them through an abstraction layer to the brain to form a concept. Our conceptual framework is simple: the agent's body carries out actions with no understanding of what they represent. Concurrently the 'mind' is processing what the body is doing linking to a corresponding table of abstractions that map to real world objects and events. To model a few of these actions and observations would indicate a sort of understanding as outlined by Jin and Bell (2003).

[1] School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, 18 Malone Rd, Belfast, BT9 5BN
Email: {**cwhite06**, **da.bell**}**@qub.ac.uk**.

Jin and Bell's account suggests that a level of understanding can be achieved by symbolic grounding. The functioning of the mind is separated into functions carried out by an actor and an observer, respectively. The actor pursues a project expressing its mood while engaging in activities. The observer records this scenario and links the abstract viewing to its own symbolic grounding. For example, the actor sporadically enters rooms while leaving a colour trail. This colour trail changes depending on what activity the actor is engaging in and if the actor is interrupted. The observer has symbolic grounding in that it can link colours to moods and areas to activities. While the actor is following a predefined path, the observer is able to reason through a single layer of abstraction, what the actor is doing and experiencing in terms of mood and activity.

3 THE ENVIRONMENT

The physical, phenomenal environment consists of a working space with 3 'rooms'. The agent can pick up attributes of the environment via sensors and with the help of a built-in observer, answer factual (sense data values) and explanatory questions about its experiences. Each activity is sought depending on what mood the agent is in. Each activity also has the ability to change the actor's mood.

The rooms are coloured differently and represent 3 distinct activities an actor can engage in (Fig.1). The agent that will be used in the final demonstration is the Khepera robot platform, and in the present preliminary investigation this was modelled by the software application Webots (Fig.2.).

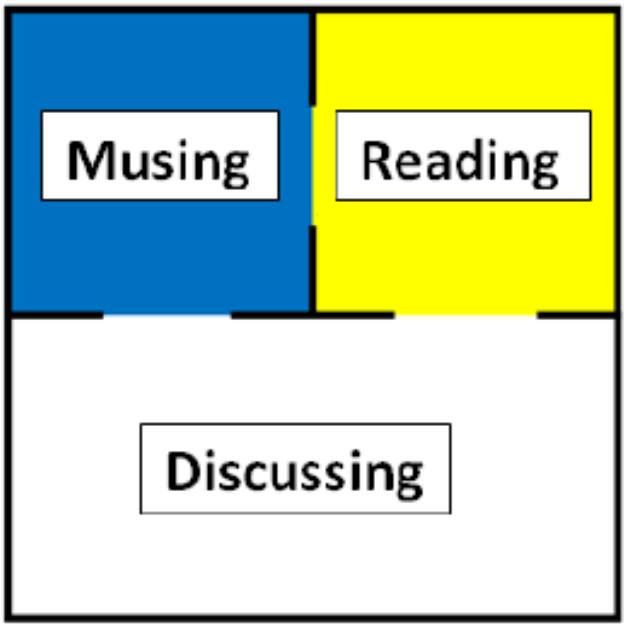

Figure 1. Overhead view of the environment, the rooms and the activities each room represents

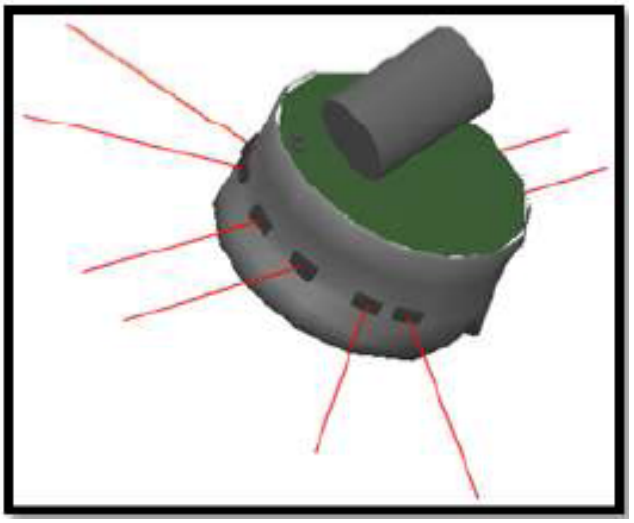

Figure 2. The Khepera agent as modelled by the Webots platform

4 THE EXPERIMENTAL DESIGN AND IMPLEMENTATION

The experiment aims to illustrate a design built on Jin and Bell's account of understanding [6]. The actor (Khepera) will follow a predefined path while leaving a colour trail based on its activity, mood and interrupts. The actor's predefined path changes with activity and interrupt. This means that if it is in a good mood, the actor will seek to read, however, if interrupted before the actor can get to the reading area, the mood and therefore the goal will change as detailed in Table 1.

The colour trail is then analysed and the pixel colour is fed to the observer. The observer can consult its memory to link the colour with a mood as shown in Table 1. The same methodology is applied to activity reasoning. The actor can constantly feed its physical location via GPS coordinates to the observer without ever understanding what it means. The observer takes this data and, again through grounding in its memory, is able to link certain boundaries with different activities as shown in Table 2. This is achieved through grounding on the observer's part; GPS coordinates of the actor as well as knowledge of room boundaries.

Finally an extra dimension is added to the experiment. We introduce the role of a 'tester' of this simple world. The tester has the ability with the touch of a button to invoke the interrupt on the actor as detailed earlier.

Table 1. How interrupts affect colour trace (mood) and the goal of the actor

Initial		After Interrupt	
Colour	**Goal**	**Colour**	**Goal**
Green	Reading	Grey	Musing
Grey	Musing	Red	Discussing
Red	Discussing	Green	Reading

Table 2. The mood data that is symbolically grounded in the agents 'Observer' memory. For example, the observer can always conclude a bad mood by the abstract link to the colour red.

Initial Colour Trail	Abstract Link
Green	Good Mood
Grey	Fair Mood
Red	Bad Mood

Table 3. The GPS data that is symbolically grounded in the agents 'Observer' memory. For example, the observer can always conclude an agent is reading when in the specified geographical area.

Room Coordinates	Abstract Link
$(0 \leq x \leq 0.45)$ AND $(0 \leq y \leq 0.45)$	Reading
$(-0.45 \leq x \leq 0)$ AND $(0 \leq y \leq 0.45)$	Musing
$(-0.45 \leq x \leq 0.45)$ AND $(-0.45 \leq y \leq 0)$	Discussing

5 Goals

The goals of this experiment are as follows:

1. Investigate the feasibility of representing Jin and Bell's simplified theory of understanding in a suitable practical platform.
2. Having successfully implemented a framework for this theory is it possible to simulate agent reasoning on a basic level? This aspect gives a two-fold question in the preliminary investigation:
    a. Can the observer successfully link colour to the mood of an actor and thus tell the tester what mood the actor is in?
    b. Can the observer successfully link physical location to GPS coordinates sent by the actor and thus tell the tester what activity the agent is engaging in?
3. Given this scenario, is there enough richness in the groundwork that has been laid to further develop this scenario? For example, can the observer reason on how an interrupt affects the mood and activity goals sought. Can the observer anticipate how an activity will affect the mood of the actor and intervene?

6 RESULTS

Results have been promising. The scenario outlined by Jin and Bell has been created approximately on the Webots platform (Fig. 3.). A trace has been successfully set up to show the actors colour trail (Fig. 4.) and a basic memory bank has been created to help the observer link, through abstraction, moods and activities based on colours and coordinates. A basic graphical user interface has been created to allow the tester to interrupt the actor at will and also ask the observer questions (Fig. 5.).

Observer correctly identified the actor's moods at each query point as well as the activity area the actor is currently engaged in. These results are preliminary in a wider research project, so it is important to assess them with regard to what basis they form for understanding and what groundwork they have laid to take this project to the next level. It is clear that by introducing classification into our analysis we can mine the output data and pose more meaningful questions to our agent. The prospect of this is very promising and it is anticipated a rich state space will be created from this data.

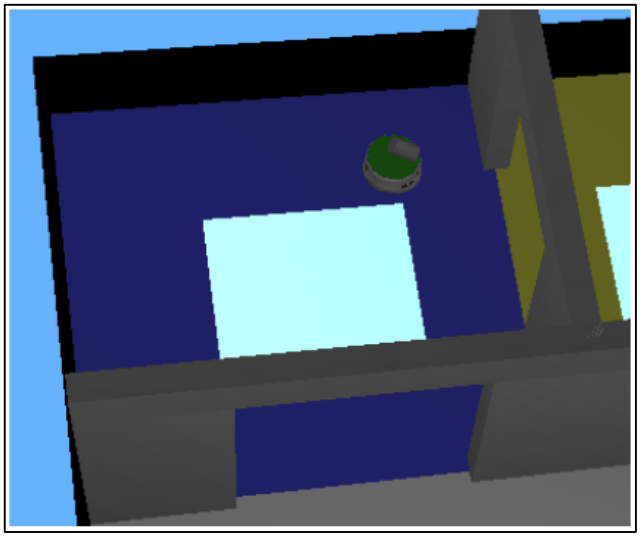

Figure 3. The agent in action on the Webots platform. The agent is currently in the musing area.

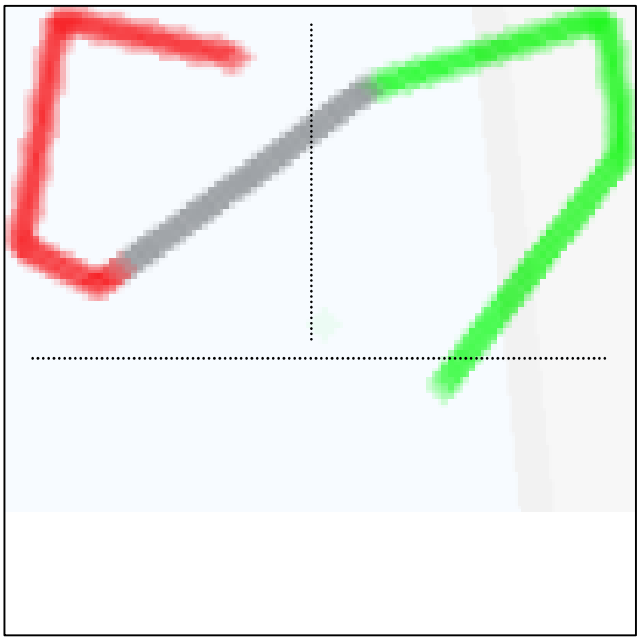

Figure 4. The trace of the agent. This trace shows the agent initially starting in the discussing area with a good mood, as the agent moves into the reading area, mood does not change. Mood changes as the agent moves to the musing area to a fair mood and then to a bad mood signified by the red colour.

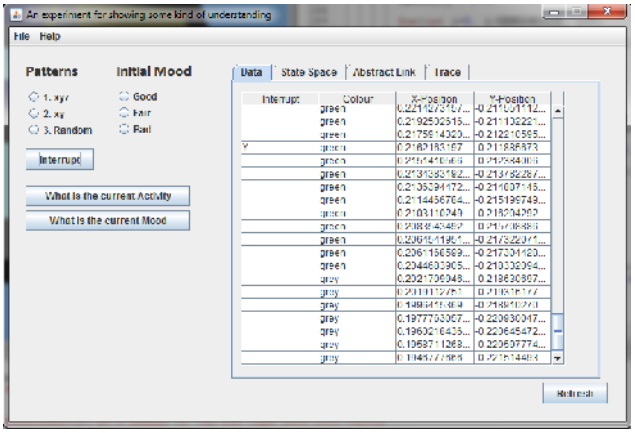

Figure 5. The basic user interface detailing the tester's options to interrupt and ask questions. Also shown is the data that the actor feeds to the observer for recording in its memory.

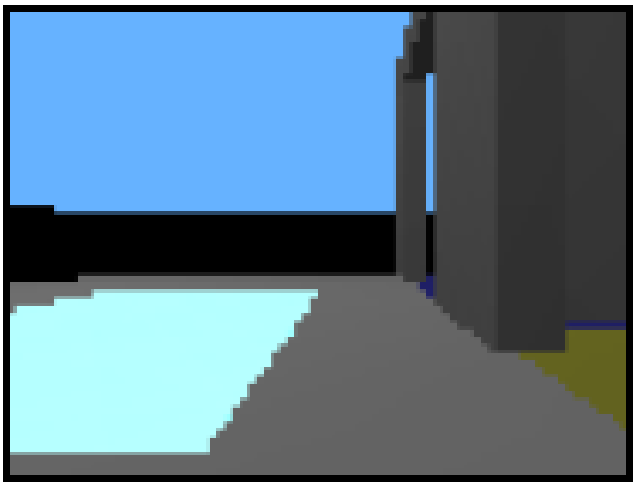

Figure 6. The view from the camera on board Khepera. While not in use in this experiment, it is clear to see that future experiments could reap the rewards of having a real 'observer' on board.

7 CONCLUSIONS & FUTURE WORK

It is hoped to extend the user interface for the 'tester' to ask deeper questions such as 'how do you know you are reading' and 'why do you want to read'. It is intended that the observer will draw from its large memory bank and make links back to the abstract data the actor had passed on as in previous straightforward questions while also incorporating rough sets and classification to create associations and rules which will help form a state space.

A camera has been attached to the agent (Fig. 6.) but is defunct in this preliminary experiment. It is hoped that utilising this camera will give a true meaning to the 'observation' taking place. This will merge two important fields of artificial intelligence (Image processing and robotics). Early ideas include determining when the actor is engaging in an activity based on imagery rather than GPS coordinates and relaying the captured images back to the observer to make sense of it. The observer will of course have still images in its memory bank to compare with.

There is a big emphasis to recreate the initial state model that the actor was defined to take part in. By building the model up and including how activities and interrupts affect moods, it is believed that this will create understanding on a basic level by linking back observational reasoning to exact predestination.

One final thought with regard to the direction of future research is the interesting topic of reinforcement learning. To have an agent reason similar to this experiment while displaying elevated examples of reinforced learning could be of prime interest in furthering understanding in autonomous agents as well as learning methods. An agent that explains to its tester why it has chosen a certain path through reasoning as well as understanding why it has been rewarded or punished is an exciting prospect. If one were to throw in the ability to affect mood similar to this experiment, it would be interesting to see the results!

REFERENCES

[1] Durkheim E, Mauss M (1903),' Primitive Classification'
[2] Weng J (2005),Muddy tasks and the necessity of autonomous mental development, in Proc 2005 AAAI Spring Symposium Series, Developmental Robotics Symposium, Stanford University.
[3] Hameroff SR (1994), Quantum Coherence in Microtubules: a Neural Basis for Emergent Consciousness, Journal of Consciousness Studies, 1, pp. 91-118.
[4] Harnad S (2003) To Cognize is to Categorize: Cognition is Categorization, ProcUQàM Summer Institute in Cognitive Categorisation, http://www.unites.uqam.ca/sccog/liens/program.html
[5] Harnad S (1990), The Symbol Grounding Problem, Physica D 42:335-346
[6] Jin Z, Bell D (2003), An experiment for showing some kind of artificial understanding, Expert Systems, Volume 20 Issue 2, Pages 100 – 107.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Towards a Staged Developmental Intelligence Test for Machines

Ed Keedwell
College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF {E.C.Keedwell@ex.ac.uk}

Abstract. The Turing Test provides a test for determining machine intelligence that has proven to be very difficult to overcome for researchers in AI, perhaps because of its pass/fail nature. A new test is proposed here, known as the Staged Developmental Machine Intelligence test which uses a number of stages based on the testing of development in children as its basis. It is proposed that this test would reduce the need for sophisticated natural language processing and provide a framework for evaluating machines on a scale rather than providing a binary result as with the original Turing Test.

1 INTRODUCTION

Some 50 years ago, Alan Turing postulated a test of machine intelligence, the Imitation Game [7] that has been generally been accepted as the benchmark that machines will be required to achieve before they can be considered intelligent. The test, now widely known as the Turing Test is in essence very simple and has been extensively described in the literature so a short description will suffice here. A human arbiter uses some form of communication (often this is described as a computer terminal) to speak to two individuals that are located elsewhere. The original task of each individual in the Imitation Game is to convince the arbiter that they are female, although further interpretations of the test consider that each individual attempts to convince the arbiter that the other is a machine. If the machine successfully convinces the arbiter that it is human, then it has passed the Turing Test and can be said to be intelligent. An important factor in this test is that the human is never in direct contact with the other individuals and so cannot make inferences based on appearance or other subconscious prejudices.

Replies to the Turing Test

As might be expected with a test of such high profile, there have been numerous replies and criticisms of this test. One of the most famous is that of the Chinese Room Argument posed by John Searle [8] in 1980. Searle states that the appearance of intelligent behaviour as required by the Turing Test is not sufficient for the machine to be considered intelligent as it is simply manipulating symbols and is not able to attach meaning to these symbols – the so called *symbol grounding problem*. A number of responses followed and Searle replied to each of these responses with largely the same notion that the symbols, whether learnt or pre-programmed, representing language, images or neural network weights, are all essentially still symbols and suffer from the same symbol grounding problem. The test proposed here is less reliant on the processing of symbolic information and so may be less prone to the arguments of Searle than the original test, although elements of the Robot Reply will likely still hold.

Issues with the Turing Test

It is proposed here that the main issue with the Turing Test is that it has a binary outcome. A system can either pass the test and be considered intelligent or fail it, with no possibility for a system to pass a portion of the test. Recent implementations of the test such as the Loebner Prize[1] have necessarily awarded a degree of 'intelligence' to the submissions as no machines has yet passed the Turing Test. The approach is to use a number of judges and a ranking system, where the winner of the prize is the system which convinces the most judges that they are most human. In short, the binary nature of the Turing Test might seem to be advantageous when considering if a system is intelligent, but as shown by the modifications utilised by the Loebner prize, this idea does not fit very well with the way that humans this about intelligence.

Degrees of Intelligence

In language, we humans often describe animals, adult humans and children as possessing a certain level of intelligence, effectively expressing intelligence as a scalar value. In particular, we discuss the level of intelligence in at least three different ways:

1. **The level of intelligence possessed by lower organisms.** This point can be characterised as a series of questions: Is a plant intelligent? Not by many standards, but it displays many more characteristics of intelligence (e.g. the ability to align itself to stimuli) than inanimate objects, a point Turing makes in his original paper. What about bacteria? Many of these have flagella or other methods of propulsion and so can seek out food or move away from danger. What about cats and dogs? Well perhaps obviously they are closer to the sort of intelligence we mean when testing machines for 'intelligence' as they can recognise patterns, have memory, can be conditioned to stimuli etc... Behaviour that we recognise as some of the more basic functions of our own intellect.
2. **The level of intelligence possessed by adult humans.** A key actor in the Turing Test is the human arbiter who makes the decision as to whether the machine is intelligent. However, surely it would be easier to convince some individuals than others? Of course we often discuss the level of intelligence or otherwise of individuals within a community, and have developed our own scalar methods of

[1] http://www.loebner.net/Prizef/loebner-prize.html

assessing intelligence through IQ tests. These tests have somewhat of a chequered past which is not within the scope of this argument, but it is clear that intelligence is considered to be a scalar property of a human, not a binary measure.

3. **The level of intelligence possessed by developing children.** The development of intelligent behaviour takes some considerable time in human children, especially the development of language. Piaget [5] postulated a four stage process of learning in his constructivist framework for cognitive development. Others have since expressed this in terms of a continuum, and whichever of these models holds, it is clear that there are some cognitive functions that have to be learned before others can then be learned. Implicit in these theories is the notion that a child does not suddenly become intelligent, but that there is a gradual building up of the faculties necessary for intelligence, again hinting at a scalar property.

These three familiar examples hopefully convince that we as humans do not consider intelligence to be a binary property, even among the adults of our own species. So if we accept that there is variation in the intelligence of our own adult population, why do we persist with a binary test for machines? A more empirical test should surely relate to the ways in which we judge ourselves and other organisms on their intelligence. In short, a Staged Developmental Intelligence Test. Aside from being aligned more closely to our own descriptions of organisms possessing levels of intelligence, a Staged Developmental Machine Intelligence Test would have the added benefit that we would be able to compare the intelligence of machines without resorting to the rather subjective ranking and repeated measures currently used with implementations of the Turing Test.

Developmental Alternatives to the Turing Test

Many alternatives have been proposed to the test since Turing originally proposed his original test in 1950 (e.g. the text compression test [4]). I will not attempt to discuss them all here, and readers are directed to [3] for a review of alternatives to the test. However, there are a handful of these tests which are relevant here. Firstly, in [6] the authors propose a test based on the acquisition of language and relate the complexity of the acquired vocabulary to the intelligence of the machine. This approach establishes the key aspect of a developmental intelligence test, the notion that machines can possess degrees of intelligence. However, the test itself is still highly dependent on the language elements of intelligence and therefore the field of natural language processing. An alternative test, known as the 'Toddler Turing Test' in [1] describes a test similar in nature to that proposed here, but only takes into account one stage of development as proposed by Piaget. The approach described here extends the idea of [1] to further stages of cognitive development without the reliance on text processing of the test shown in [6].

2 COGNITIVE DEVELOPMENT

The ideas of learning and cognitive development are closely associated with the development of mature intelligence that might be tested by the Turing Test. Many theories exist as to how human infants learn about the world but perhaps the most widely known are those of Piaget. Piaget's constructivist view [5] of development contrasted starkly with the opposing behaviourist and nativist approaches that existed at the time. His ideas were among the first to suggest that both nature and nurture play a part in the cognitive development of children. Although more recent theories have superseded Piaget's work somewhat, his ideas remain highly influential in many areas of education including education policy. The principles of testing at 7 and 11 years of age are as result of his staged approach to development. More recent theories (notably Siegler [9][10]) have shed some doubt as to the validity of a staged approach, suggesting that there is overlap between the stages, but the Staged Development Intelligence Test suggested here could be adapted to new development schemes as they are put forward.

Piaget's Stages of Cognitive Development

Piaget's theory [5] states that cognitive development occurs in four stages, during which a child learns new concepts. Importantly, if a child is in one stage, the theory states that they cannot learn concepts from another, more advanced stage. The four stages are shown in Figure 1:

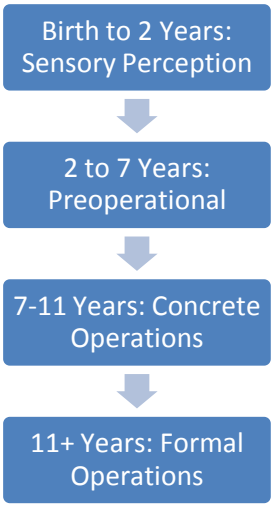

Figure 1 - Piaget's Stages of Cognitive Development

Sensory Perception

In this stage infants are able to distinguish themselves from others and begin to act intentionally on their environment and are able to begin to recognise object permanence. However the actions available are very much dictated by the current stimulus, infants in this stage for instance will smile at a parent's face but cannot intentionally recall that image and smile again.

Pre-Operational

In this stage, children are able to use language to represent objects and to classify objects into classes according to common attributes (e.g. colour). However children in this stage still find impossible to consider other people's points of view and are therefore highly egocentric.

Concrete Operational

Children can think logically about objects and events and can begin to conserve properties of objects (i.e. number (age 6), mass

(age 7) and weight (age 9). Also more advanced classification and sorting behaviours are demonstrated in this stage.

Formal Operational
At this stage children can apply logic to abstract events and objects and can test various hypotheses about the world. The child can also consider hypothetical scenarios, including those that might occur in the future.

It is clear from this framework that full adult intelligent reasoning is built from the ground up. Certain characteristics of the world must be learned before other, more sophisticated concepts can be considered. This is perhaps most clearly demonstrated by the idea of objects – firstly their permanence in the world must be established and then they can be classified according to certain properties they possess. Some of these properties are then learned to be conserved, no matter what transformation is applied to them. Finally objects can be considered in abstract situations and the consequences of actions upon them can be visualised rather than needing to be carried out.

3 IMPLICATIONS FOR THE TURING TEST

This developmental framework above shows that the original Turing Test, due largely to its generality, tests all of the stages of cognitive development (and therefore what we understand as intelligent behaviour), up to and including formal operations. The sorts of questions asked by judges of the Turing Test tend to be experiential in nature, so questions such as "where are you from?" and "have you any plans for later in the day?" are commonplace in the transcripts of competitions such as the Loebner prize.
So whereas we might reasonably expect fully grown adults to be able to answer such questions competently, we wouldn't expect 2 or 3 year olds to engage in conversations that involve these higher concepts of abstract thought. The concepts are quite abstract in nature and therefore test the latter stages of the developmental framework as shown above.

Certainly, as a gold-standard test of whether machine intelligence exists, the Turing Test provides us with a simple-to-implement and robust evaluation of machine intelligence. However, the test itself is not very helpful in the development of a machine intelligence that might one day be able to pass it. By posing such a wide and apparently insurmountable problem to researchers, perhaps we are encouraging the sort of 'tuned' approaches and flawed responses that Turing sought to reduce. The contention here is that we should not be rewarding AI approaches that get marginally closer to the prize by becoming infinitesimally more able to respond in somewhat more human-like terms without materially improving on the intelligence behind the approach.

4 A STAGED DEVELOPMENT TURING TEST

A test of intelligence should be one that measures machine intelligence in broadly the same way as we measure a child's, through a staged approach to cognitive development. Suggestions for the tests are given below. In each case, the human behaviour is first described and is followed by a possible machine test to determine whether the machine has reached this level of capability. The tests are intended to be as generic as possible, but it is perhaps easiest to see how this might work best with a system that processes visual information and can make changes to the environment through robotic (or simulated robotic) action.

Stage 1 – Sensory Perception Stage
In this stage the test would initially involve simple stimulus-response tasks that are conducted with infants. There are a number of these used to assess progression from Birth through to 2 years.

Reacts to basic stimuli – light/sudden sound
- The machine is able to 'attend' to important stimuli (e.g. the facial recognition systems available in some cameras).
- The system will attend to sharp changes in the stimulus.

Understands cause and effect
- The machine can predict the likely movement of objects when the video stimulus is stopped (e.g. a bouncing ball hanging in mid-air). The accuracy of this prediction can be used to determine the degree of proficiency of the machine.

Understands the concept of objects and what to expect from them
- The machine can differentiate between objects of different types (essentially a classification task). In particular, those objects that it can directly effect (e.g. move) and those that it cannot.

Uses trial-and-error to learn about the world
- The machine possesses the ability to modify its internal structure based on experimentation with the world.

Stage 2 – Pre-Operational
The development of language skills comes to the fore in this testing stage. Systems that would pass this stage would be able to demonstrate:

Language understanding relating to itself (but not wider topics)
- The machine should be able to answer questions about itself and the current state it is in. Although still a sophisticated NLP task, the reduction in complexity over the standard Turing Test should make this restricted goal more achievable.

Can relate to objects through language even though they are not present in the perceptual field of the system
- The machine would be required to remember (requiring memory) information about objects that it has 'seen' previously.

Stage 3 – Concrete Operations
Higher order thinking starts to become evident in this phase. Systems that would pass this phase would demonstrate
Conservation of volume, mass etc..
- In this test, the machine should be able to determine the correct quantity of objects despite their arrangement in the visual stimulus.

Classification of objects based on logical basis rather than superficial object characteristics
- The machine should be able to separate objects into logical groups (e.g. animals, shapes etc..)

Sorting of objects

- Given a set of objects, the machine should be able to order them according to some criterion (e.g. size or colour).

Reversibility

- When shown an action taking place (e.g. placing a weight on some scales), the machine should be able to accurately predict what would happen if that action were reversed (e.g. the weight was removed).

Stage 4 – Formal Operations

Logical thinking becomes evident in this phase. Systems that would pass this phase would demonstrate:

The ability to create its own hypotheses

- The machine should no longer be directed by the learning process but should be able to conduct its own experiments to test those hypotheses. (Machines already exist that are able to conduct this hypothesis testing in the restricted field of science [2]).

Abstract Thought

- The machine should be capable of considering stimuli not presented to the machine and should be able to consider the interactions of objects, seen or unseen, in novel ways.

Final Stage

A final stage could be implemented whereby the existing Turing Test (or a visual version of the same using objects rather than language) could be used to ultimately determine that an intelligent machine has been created.

Machines can then be judged according to their effective 'stage' from the framework above. Machines that are capable of all concepts within a stage could then win that particular stage and a new competition would be opened for the next stage. Further agreement and standardisation of the tests would need to be undertaken before it would be ready to be developed as a test in its own right and the test could be adapted for other theories of cognitive development (e.g. Siegler's overlapping waves[9][10]) if required. However, it is the principle of a staged approach is the important contribution here.

A further consideration is that it may be that even the capabilities of the first stages would be too complex for current machines and therefore a finer-grained test could be conducted which adheres to a well known developmental tests (e.g. "Schedule of growing skills", Denver 2, Griffiths and Mary Sheridan developmental tests) which typically test children to a maximum of age 8. These tests all work within the first two stages of Piaget's framework and so would provide a more detailed rating system for machines in these stages.

5 DISCUSSION

This developmental framework shown here has a number of advantages. Firstly, the stage reduces the focus on natural-language processing, a very difficult field of AI which is often the stumbling block for machines wishing to pass the Turing Test and restricts the AI approaches that can be taken to passing the Turing Test to those that can efficiently process symbols. This could lead to more machines (e.g. Asimo) which focus on the perceptual aspects of intelligence rather than the language-based elements, particularly if the visual stimulus tests were conducted as shown in section 4.

Secondly, aspects of intelligence would be investigated in order, leading to a developmentally focussed approach to delivering machine intelligence. For instance object permanence and the ability to manipulate objects would come at an early stage in the process and would be later built upon. This developmental aspect means that we would create machines that more closely mimic our own progress towards intelligence rather than trying to create an intelligent machine from scratch. We tend to associate intelligence with the final product of development, i.e. the adult human, but the process of development may be crucial in the creation of intelligent machines.

Thirdly this test provides an established framework within which to classify machines as to how 'intelligent' they are. Currently machines can only be classified as passing or failing the Turing Test, of which only machines in the latter category exist at present. The impasse that has existed for 60 years would be lifted somewhat as progress would be measurable, even though the ultimate goal remains the same.

However, there will be criticisms of this approach and in the spirit of Turing's original paper some possible replies to the Staged Approach are presented here. Firstly, more than ever, this test would doom us to creating intelligence in our own image. This could certainly be interpreted as a negative aspect as replicating human intelligence processes in a machine might limit the possibilities of what can be achieved with silicon. To counter this I would point to the fact that currently we cannot come anywhere close to replicating our own intelligence, let alone create a new species of intelligent behaviour. A pragmatic way forward would be to create thinking machines in our own image that can pass all tests and only then consider how we might allow the machines to develop differently. Quite simply, if our own intelligence is all we have as a basis for intelligent machines, then shouldn't replicating that be our first priority? In any case, there is no guarantee that creating a human developmental process within a computing machine with its vastly different architecture and capabilities would inevitably lead to human-like intelligence.

Secondly, a related reply might be that the method of achieving intelligence should be independent of any test, perhaps an intelligent system might exist that does not need to develop in the stages shown here. My response to this is that a key point of this test is that it is based on the behaviour of the system in a similar way to the Turing Test and as such is independent of the implementation used. Also it does not prescribe that the system should necessarily develop in this way, but crucially we would expect a system that was capable of passing the final stage would also be capable of passing the previous stages, in the same way that one would expect a 11 year old child to be able to count and conserve volume as well as be capable of abstract, logical thought.

6 CONCLUSION

A Staged Development Intelligence Test has been presented with the main argument that if we are to judge intelligent machines by our own standards of intelligence, then we can enrich the process of AI development by considering degrees of intelligence

displayed during cognitive development. The stages can be considered as milestones towards ultimately solving the original Turing Test, or similar variant, as the final stage in the development of thinking machines.

7 ACKNOWLEDGEMENTS

Thanks to Lynda Keedwell for some informative discussions on child development.

8 REFERENCES

[1] Alvarado, N., Adams, S., Burbeck, S., Latta, C., (2002) "Beyond the Turing Test: Performance Metrics for Evaluating a Computer Simulation of the Human Mind" in *Performance Metrics for Intelligent Systems2002*

[2] King *et al.* (2009) "The Automation of Science" *Science* Vol. **324,** No. 5923, pp 85-89

[3] Legg, S., Hutter, M. (2007) "Tests of Machine Intelligence" in *50 Years of Computer Science LNCS* **4850**, pp232-242

[4] Mahoney, M (1999) "Text Compression as a Test for Artificial Intelligence" in *Proceedings of AAAI/IAAI*

[5] Piaget, J., (1964) "Cognitive Development in Children", *Journal of Research in Science Teaching*, **2**, No 3., pp176-186

[6] Triesten-Goran, A., Dunietz, J., Hutchens, J.L., (2000) "The developmental approach to evaluating artificial intelligence" in *Performance Metrics for Intelligent Systems 2000*

[7] Turing A., (1950) "Computing Machinery and Intelligence" *Mind* **59**, pp433-460

[8] Searle, J., (1980) "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, 417-424.

[9] Siegler, R.S., (2000) "The Rebirth of Children's Learning" *Child Development*, **71**, No 1., pp26-35

[10] Shrager, J., Siegler. R.S., (1998) "SCADS: A Model of Children's Strategy Choices and Strategy Discoveries" *Psychological Science*, **9**, pp405-410

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

How to Detect an Android

Antoni Diller[1]

Abstract. A number of prominent researchers in Artificial Intelligence look forward to a time when androids will co-exist with humans. In such a world the question 'Can machines think?' will be answered in the affirmative, if it is asked at all. A far more pressing problem will be whether androids could be differentiated from humans in some way. It is argued that an undetectable android cannot be manufactured and a method is presented for distinguishing between androids and humans.

1 INTRODUCTION

Turing [22] famously devised a test, the imitation game, whose purpose was to investigate the question whether human intellectual ability was significantly different from that of a machine. The nature of the test was influenced by the technology of the day. An interrogator and two test subjects, one male and one female, occupy different rooms. The interrogator can be either a man or a woman. They communicate by means of a teleprinter. The interrogator has to determine which of the subjects is male and which female. He does this by asking both of them a series of questions. The man is instructed to try and fool the interrogator into making an incorrect identification, whereas the woman is told to aid the interrogator in his or her task. The game is played in this form several times. The male subject is then replaced by a computer. Presumably, the interrogator knows this has taken place. (Unfortunately, Turing's account of the imitation game is quite sketchy.) It is now the computer's task to fool the interrogator into making the wrong identification. Again, the game is played in this form several times. Turing [22, p. 434] replaces the question 'Can machines think?' with: 'Will the interrogator decide wrongly as often when the game is played like this [that is, between a machine and a woman] as he [or she] does when the game is played between a man and a woman?' (Turing thought a time would come when he or she would.) This question presupposes the game is played several times in both versions. It is not clear, however, if the participants are different each time. I think they would have to be for the procedure to make sense, but as the unearthing of Turing's precise meaning is not my main concern here, I will not pursue the matter.

Inspired by Turing's work, Hugh Loebner created the Loebner Prize in the early 1990s. He offered a sum of $100,000 and a gold medal to the first person who could devise a program that could fool ten judges into thinking it was human on the basis of a conversation, about anything, lasting three hours. This prize has still to be won. A prize for the most human-like program is, however, awarded each year. Programs in this competition are only expected to converse about a single topic for just a few minutes. Reading through transcripts of entrants' dialogues shows that the programs are becoming increasingly sophisticated and that the gold medal may, indeed, be won one day. (These transcripts, and further information about the prize, are available on the `www.loebner.net` website.)

Technology has improved dramatically since Turing's day and many workers in Artificial Intelligence (AI) and Robotics today are far more ambitious than he ever was. They would agree with Brooks and his colleagues when they write: 'Building an android, an autonomous robot with humanoid form and human-like abilities, has been both a recurring theme in science fiction and a "Holy Grail" for the Artificial Intelligence community' [3, p. 52]. Considerable progress has been made in achieving the goal of manufacturing an android and one day it is likely that very life-like machines will co-exist with humans. In such a world the issue will no longer be whether machines can think, but how to distinguish between humans and androids.

2 NEW QUESTIONS

Is it possible to construct a robot that looks like a human being and which can live in human society and be taken for a human being? Is it possible for a machine made out of mechanical and electronic components to behave in exactly the same way as a human? Some scientists working in AI and Robotics believe that it is. Brooks thinks that it is just a matter of time before an android is built that can pass for human [2, p. 209]. Moravec believes this will happen by 2040 [19, p. 88]. In this paper, I argue that it is impossible for human beings to design and build an undetectable android. In particular, I show that human emotion-related behaviour will always be distinguishable from android emotion-related behaviour. I focus on emotion-related behaviour in order to make my reasoning easier to follow. With suitable changes, however, the argument could be used to show that any clearly distinguishable type of human behaviour could never be perfectly replicated by an android. After showing that androids will always be detectable, I present a method that could be used to identify androids once they start living amongst us.

3 DESIGNING ANDROIDS

Before any complex mechanism can be built, it first has to be carefully designed. The design team has to make use of a large number of theories. This applies as much to the design of an android as it does, say, to that of an aeroplane. When crafting an aeroplane, the design team needs experts in many fields, including aerodynamics, ballistics, chemistry, mechanics, metallurgy and structural engineering. Similarly, when fashioning an android, the design team will comprise people with expertise in several disciplines. They will require knowledge of theories concerning many different things. Some of these will be physical theories relating to the android's mechanical components. The android's limbs, for example, may be moved using hydraulic pumps and so some members of the design team will need to have a thorough knowledge of hydraulics. Some of the theories,

[1] University of Birmingham, UK, email: A.R.Diller@cs.bham.ac.uk

however, will relate to the android's social behaviour. Knowledge of physics will enable the design team to produce an android that can shake hands, say, but not one that knows when it is socially acceptable to do so. I am interested in those theories that relate to the android's social behaviour and intellectual abilities. Some psychologists use the term 'social brain' to refer to that aspect of a human being responsible for producing emotions and emotion-related behaviour. This is a convenient label, but by using it I do not wish to suggest that I believe there is an identifiable region of the physical brain that is responsible for everything to do with emotions in human beings. In this paper, I am particularly interested in the problem of designing an android's social brain.

It should be noted that, using a distinction introduced by Dennett [5, p. 194], the android would *instantiate* the theories used in its design and not merely *incorporate* them. A computer simulation of a hurricane, say, incorporates a theory of hurricane behaviour. It is a computer 'program, which, when you feed in *descriptions* of new meteorological conditions, gives you back *descriptions* of subsequent hurricane' behaviour (p. 191); it does not produce such behaviour. The android I am considering, however, would instantiate various theories, because its behaviour would be generated by those theories; they would not produce descriptions of human behaviour.

In order to live in a human society and interact meaningfully with human beings an android must be able to display and recognise human emotions. Currently, our understanding of the nature of emotion is lamentable. There are many different theories of emotion competing for our attention. For example, Strongman [21], in chapter 2 of his book *The Psychology of Emotion*, discusses about thirty different theories, but he does not claim to have mentioned every theory there is. He admits [21, p. 13]: 'To describe all theories of emotion would necessitate a book in itself, so the number has been restricted.' The fact that there are so many competing and incompatible theories strongly suggests that none of them are very good, for the existence of a good theory, or a small number of good theories, would quickly lead to its competitors being consigned to the dustbin of history. So, using an existing theory of emotion in the design of an android's social brain is highly unlikely to result in the android exhibiting emotion-related behaviour which is indistinguishable from that of a human. However, even if our understanding of emotion improved dramatically, that would not increase the chances of making an undetectable android, because, no matter how good our theories of emotion become, they will all always be false. This follows from the fact that all scientific theories are false. (It should be noted that this is very different from saying that they are all useless and worthless, as I explain below.) This claim is an essential premise in my argument to show that there cannot be an undetectable android. To many people this claim seems absurd when they first come across it, but a little investigation shows it is much more widely accepted than you would initially think. Lakatos, for example, liked to say that all scientific theories are born refuted, live in an 'ocean of anomalies' and die refuted [16, pp. 5, 126, 128 and 147]. Furthermore, there is much discussion, by philosophers and historians of science, of an argument known as the *pessimistic induction* (or *meta-induction*). This starts from the observation that the history of science is littered with discredited theories. These include the theory of spontaneous generation, the caloric theory of heat, the theory of global cooling (widely accepted in the 1970s), Aristotelian mechanics, Ptolemaic astronomy, Kepler's celestial mechanics, geological catastrophism, the medical theory of humours, the effluvial theory of static electricity, the theory that the chemical atom is indivisible and the phlogiston theory. This list could easily be extended with many more examples. Newton-Smith [20, p. 183] presents the pessimistic induction as follows:

> Past theories have turned out to be false, and since there is no good reason to make an exception in favour of our currently most cherished theories, we ought to conclude that all theories which have been or will be propounded are strictly speaking false.

He even goes so far as to say [20, p. 14]: 'Indeed the evidence might even be held to support the conclusion that no theory that will ever be discovered by the human race is strictly speaking true.' (In order to avoid paradox it is important to emphasise that the conclusion of the pessimistic induction is not itself a scientific theory of the same sort or level as the theories it applies to. Recent work on the pessimistic induction includes papers by Hobbs [15], Lewis [18], Lange [17] and Busch [4].)

Inductive arguments are not deductively valid and so it is possible for their conclusions to be false even when all their premises are true. They are, however, often useful as heuristics. In science, for example, they are frequently helpful in suggesting universal theories that the scientific community then tries to refute. The inability to falsify a theory increases our confidence in its verisimilitude. Why theories are added to or removed from the body of generally-accepted scientific knowledge is a complex matter which is still being studied by philosophers of science. The difficulty of the issue was revealed by the pioneering work of Kuhn, Lakatos, Feyerabend and Laudan. The main problem is that counter-examples can be found to the various rational criteria of acceptability that have been put forward. (Some of this research is admirable summarised in chapter 4 of Bechtel's book *The Philosophy of Science* [1], written for cognitive scientists.)

The considerations presented show that the conclusion of the pessimistic induction has not been conclusively established, but the pessimistic induction does provide strong *prima facie* evidence in its favour. Much more could be said about the view that all scientific theories are false, but my main concern is not to present an exhaustive account of the debate that the pessimistic induction has given rise to. I will just say, however, that the debate is exclusively about how good an argument the pessimistic induction is. No one, as far as I know, tries to falsify its conclusion by presenting a true and completely accurate scientific theory which is universally acknowledged as such.

If the reader is undecided about the correctness of the view that all scientific theories are false, then he or she may still be interested in how I derive the claim that an undetectable android cannot be made from this position, together with some other considerations, and especially in the following conditional sentence, which is established if my reasoning is correct: If all scientific theories are false, then an undetectable android cannot be made. Someone who is in two minds about the claim that all scientific theories are false is still likely to find the contrapositive of the previous conditional sentence downright counter-intuitive: If an undetectable android can be made, then some scientific theories are true.

Although many people think that is is reasonable to believe that all scientific theories are false, this position does not have disastrous consequences for science or technology. Just because a theory is false it does not follow that it is useless. Newton's theory, for example, is false, but it was used by NASA scientists in order to plot the trajectories of rockets sent to the moon and those rockets reached their destination. It is also regularly used by scientists in their preparations to launch satellites. Although Newton's theory has been falsified, it is still a good approximation to the truth. Furthermore, some theories which we now regard as radically mistaken, such as the medical

theory of humours and the eighteenth-century optical aether theory, were incredibly successful and useful in their day.

Returning to the question of designing the social brain of an android, let T be the theory of emotion used by the design team in order to develop and build the android's social brain C. The android's emotion-related behaviour is produced by C. Using Dennett's terminology, C instantiates T. Theory T is a model of human emotion-related behaviour, a conjecture as to how such behaviour is produced by a human's social brain. In these circumstances it would be possible for human behaviour and emotional response to falsify theory T, but it would not be possible for android behaviour and emotional response to falsify T. As a theory of human emotion, T is a falsifiable, empirical theory, whereas as a theory of android emotion, produced by component C, it is an unfalsifiable, non-empirical theory. It would be possible for theory T to be falsified by experiments involving human beings, but it would be impossible for android behaviour to falsify the theory, as that behaviour is produced by T. Because all scientific theories are false, we know that T is false and some human behaviour does actually falsify it. This means that the difference between human emotional response and android emotional response can be established by any human behaviour that falsifies theory T, because an android could not exhibit such behaviour. Thus, there is a way of detecting the presence of androids in human society.

An example may clarify the above argument. This example has been simplified, but it still accurately presents the issues involved. Consider a situation in which an android has been equipped with a cognitive theory of emotion as developed, for example, by Dryden [14]. The specific emotion that I shall discuss is anxiety. In such a theory people feel anxious and exhibit anxiety-related behaviour if they have a cluster of irrational beliefs about some imminent, personally significant event. For example, imagine someone facing a job interview and holding the irrational beliefs: 'I must get this job', 'It would be terrible if I failed to get this job', 'I couldn't stand not getting this job' and 'If I fail to get this job, that proves I'm worthless and that I'll never get a decent job'. According to the cognitive theory such a person would feel extremely anxious and exhibit some anxiety-related symptoms. For the sake of argument, I take these symptoms to be disturbed sleep, dryness in the mouth, trembling and sweating. An android equipped with this theory of emotion, holding these irrational beliefs and facing a job interview would exhibit the anxiety-related symptoms listed above. Moreover, whenever it was going for a job interview having these irrational beliefs, it would exhibit the same symptoms. In the case of a human being, however, it is not logically impossible to conceive of a person holding these irrational beliefs, while waiting for a job interview, and yet not exhibiting some of these symptoms. In fact, because of the plausibility of the claim that all scientific theories are false, we can be confident that the cognitive emotion, no matter how useful it is in psychotherapy, is false and some human behaviour does actually falsify it. For the sake of argument, let us assume that a person with the above irrational beliefs facing a job interview does not exhibit any trembling. An android in whom the cognitive theory was instantiated would tremble, however, in these circumstances. Thus, the presence of trembling in an entity looking like a human being would alert us to the fact that we were dealing with an android.

Perhaps an analogy will make my reasoning clearer. An orrery is a clockwork model of the solar system. An orrery could be built to illustrate Ptolemy's celestial mechanics. The Sun and planets would move around the Earth in orbits produced by combining several circular motions. Another orrery could be made to illustrate Copernicus's theory in which the orbits of the planets are again obtained by combining several circular motions, but now the Sun is at the centre of the system. Yet another orrery could be fashioned to illustrate Kepler's celestial mechanics. In this the planets would move in ellipses around the Sun which would lie at one of the foci of each of these ellipses. In reality, an orrery could not be built which is an exact scale model of the solar system, because the highest common factor of the mean distances of the planets from one another is minute in comparison with the mean distance of the furthest planet from the Sun. A computer simulation could, however, be fashioned. Imagine such a simulation built to illustrate Kepler's celestial mechanics. The behaviour of the simulation could not deviate from that described by Kepler's theory. No matter how many readings we took of the simulation they would always be in conformity with that theory. It would be impossible for any such readings to falsify Kepler's theory. We know that the behaviour of our solar system is different from what we would expect on the basis of Kepler's theory. The behaviour of our solar system that falsified Kepler's celestial mechanics could not be produced by the computer simulation.

In this analogy, the computer simulation (or orrery) corresponds to an android and the behaviour of the simulation corresponds to the android's emotion-related behaviour, the real solar system corresponds to a human being and its behaviour corresponds to human emotion-related behaviour. Kepler's celestial mechanics corresponds to the theory of emotion used to produce the android's emotion-related behaviour. In certain circumstances, there are discrepancies between android and human behaviour, just as there are discrepancies between the behaviour of the simulation and the behaviour of the solar system. In the case of emotion-related behaviour these discrepancies allow us to differentiate between androids and humans.

4 THE METHOD OF DETECTION

The above argument shows that there must be a discernible difference between human and android emotion-related behaviour. On the basis of these considerations a method can be devised to ascertain whether a test subject is an android or a human. Let us assume that we suspect our subject to be an android manufactured by a particular corporation. We first have to find out which theory of emotion T scientists working for that corporation use in designing their androids. This theory is instantiated in the social brains of all their androids. Then, we have to find out what human behaviour B falsifies T. After that, we have to put the test subject in a situation where a human would exhibit behaviour B. If the subject exhibits B, it is human. If the subject fails to exhibit B, it is an android. As behaviour B falsifies T, it could not be exhibited by the android.

It is possible that different design teams, working for other companies, might use various theories of emotion when designing androids. It might, therefore, be necessary to use the above method several times in order to find out if a test subject is an android. It might be necessary to use the above method as many times as there are theories of emotion that have been instantiated in different ranges of android. Furthermore, as theories of emotion get more sophisticated the behaviour B that we need to discover to detect androids will get harder to find, but so long as we deal with scientific and empirical theories of emotion we can be confident that such behaviour exists.

5 CONCLUSION

Unlike some people, I am not frightened by the prospect of intelligent, autonomous androids living side-by-side with humans. Furthermore, I believe this will happen one day. I am irritated, however,

by people who say that this will happen in the near future, maybe in the next fifty years or so. This is because there are certain human intellectual abilities that hardly anybody is trying to mechanise and, until they are incorporated in androids, humanoid robots will have no chance of competing intellectually with humans. For example, the capacity to acquire information from testimony, that is, from what others say and have written, is crucial to our lives in society. An android would also need to have this aptitude in order to live, work and interact meaningfully with humans, yet virtually no one in AI and Robotics is studying how to give androids this ability. Not until we can give androids this aptitude will they have a chance of being truly intelligent, as I have been arguing for many years [9, 12, 13]. I am, however, neither a Luddite nor a Jeremiah. I have been studying testimony from an AI perspective in order to work out how to equip androids with the capacity to learn form other people's assertions [7, 8, 10, 11, 13]. It is important that androids should have this ability. Be that as it may, in this paper my purpose has been different. I have argued that it will always be possible to devise a method to tell whether a human-looking being is really human or whether it is an android. This is not meant to deter people from trying to make ever more life-like robots. I want people to continue producing ever more sophisticated androids. It is, however, the fact that they are designed and built by fallible human beings that makes them detectable.

ACKNOWLEDGEMENTS

A preliminary version of the argument presented here, written for a general readership, appeared in a simplified form in the popular magazine *Philosophy Now* [6]. I have also expounded it in several lectures, talks and seminars over the years; I am grateful for the feedback that I have received.

REFERENCES

[1] William Bechtel. *Philosophy of Science: An Overview for Cognitive Science*. Tutorial Essays in Cognitive Science (advisory editors: Donald A. Norman and Andrew Ortony). Lawrence Erlbaum Associates, Hillsdale (New Jersey), 1988.

[2] Rodney A. Brooks. *Robot: The Future of Flesh and Machines*. Allen Lane: The Penguin Group, London, 2002.

[3] Rodney A. Brooks, Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. The Cog project: Building a humanoid robot. In C. Nehaniv, editor, *Computation for Metaphors, Analogy, and Agents*, volume 1562 of *Lecture Notes in Computer Science*, pages 52–87, Heidelberg, 1999. Springer-Verlag.

[4] Jacob Busch. Entity realism meets the pessimistic meta-induction argument. *Sats: Nordic Journal of Philosophy*, 7(2):106–126, 2006.

[5] Daniel C. Dennett. Why you can't make a computer that feels pain. In *Brainstorms: Philosophical Essays on Mind and Psychology*, chapter 11, pages 190–229. Harvester Press, Brighton, 1981.

[6] Antoni Diller. Detecting androids. *Philosophy Now*, (25):26–28, Winter 1999/2000.

[7] Antoni Diller. Acquiring information from books. In Max Bramer, Alun Preece, and Frans Coenen, editors, *Research and Development in Intelligent Systems XVII: Proceedings of ES2000, the Twentieth SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, December 2000*, pages 337–348. Springer, London, 2001.

[8] Antoni Diller. A model of assertion evaluation. Cognitive Science Research Papers CSRP–02–11, School of Computer Science, University of Birmingham, November 2002.

[9] Antoni Diller. Designing androids. *Philosophy Now*, (42):28–31, July/August 2003.

[10] Antoni Diller. Modelling assertion evaluation. *AISB Quarterly*, (114):4, Autumn 2003.

[11] Antoni Diller. Assessing information heard on the radio. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining: Proceedings of the International IIS:IIPWM'05 Conference held in Gdańsk, Poland, June 13–16, 2005*, Advances in Soft Computing, pages 426–430. Springer, Berlin, 2005.

[12] Antoni Diller. How empiricism distorts AI and Robotics. In M. H. Hamza, editor, *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2005)*, pages 339–343. ACTA Press, Anaheim, Calgary, Zurich, 2005.

[13] Antoni Diller. Why AI and Robotics are going nowhere fast. In Jordi Vallverdú, editor, *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles*. IGI Global, Hershey (PA), *in press*.

[14] Windy Dryden. *Invitation to Rational-emotive Psychology*. Invitations to Psychology. Whurr, London, 1994.

[15] Jesse Hobbs. A limited defence of the pessimistic induction. *British Journal for the Philosophy of Science*, 45(1):171–191, 1994.

[16] Imre Lakatos. *The Methodology of Scientific Research Programmes: Philosophical Papers, volume 1*. Cambridge University Press, Cambridge, 1978. Edited by John Worrall and Gregory Currie.

[17] Marc Lange. Baseball, pessimistic inductions and the turnover fallacy. *Analysis*, 62(4):281–285, October 2002.

[18] Peter J. Lewis. Why the pessimistic induction is a fallacy. *Synthese*, 129:371–380, 2001.

[19] Hans Moravec. Rise of the robots. *Scientific American*, 281(6):86–93, December 1999.

[20] W. H. Newton-Smith. *The Rationality of Science*. International Library of Philosophy, editor: Ted Honderich. Routledge & Kegan Paul, Boston, London and Henley, 1981.

[21] K. T. Strongman. *The Psychology of Emotion*. John Wiley & Sons, Chichester, second edition, 1978.

[22] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, October 1950.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Don't Improve the Turing Test, Abandon It

Drew McDermott[1]

Abstract. This workshop is a bad idea, at least if held as part of an AISB meeting. When Turing suggested replacing the question, "Can a machine think?" with the question, "Can a machine pass such-and-such a test?," he never intended for the replacement to be a permanent institution, or for it to form part of a proposed definition or even inductive test for intelligence. Some philosophers may have based an entire subfield of philosophy of mind on this misunderstanding, but AI as a field need not follow them, or respond to them.

1 Intro

"The aim of this symposium is to reconsider the Turing Test ... with the goal of outlining a new test ... that may more usefully be employed to evaluate 'machine intelligence' ..." [1]. I will argue that any symposium with this aim is a bad idea, because the Turing Test is a bad idea. Although this point has been argued eloquently already by Hayes and Ford [6], I will make the case from a slightly different angle. I will argue that Turing's intentions were misunderstood, and that it is only the insecurity of researchers in cognitive science that has kept endless discussions about his proposal on the table.

2 Turing Would Have Repudiated Turing-Style Tests

Alan Turing never doubted that people were machines [9]. This view may have stemmed from the oddness of his personality; the thought that he was a machine may have come more naturally to him than to the average person [12]. From the beginning, when he thought about computational issues, even before computers actually existed as physical entities, he assumed that similar things were going on inside human brains and the computers he imagined. Initially the analogy ran from brains to machines: In his seminal paper [21] on what are now called Turing machines, in thinking about what machines could calculate, he worked from idealized models of what a person following an algorithm could calculate, breaking the process down into the simplest steps we can conceive of [21, sect. 9]. But by the 1940s, the process was going the other way. Now the physical Turing machines did exist. He had access to the Manchester machine [8, 23], and even though its computing capacity seems laughable today, he assumed, correctly, that technology would change the boundary conditions dramatically and soon. Now he wanted to know when computers would acquire abilities comparable to their creators'.

Like everyone else, he made the mistake of assuming that sensory and motor skills would be relatively straightforward to decipher, so that the difficult part of AI would be intellectual skills such as mathematics and playing chess. Chess was one area in which it was possible even in the late forties to envision competing against humans, so Turing worked on a chess program, although it was never fully implemented.

However, he was impatient with frontal attacks on the AI problem, and wanted to get to the heart of the matter quickly. Hence the speculative paper "Computing Machinery and Intelligence," which appeared in *Mind* in 1950.[2] In the first paragraphs of this paper, Turing suggests a strategy for speculation:

> I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used

[1] Computer Science Department, Yale University, New Haven, CT, USA email: drew.mcdermott@yale.edu

[2] Turing's predictions for how easy different tasks would be, and how powerful computers would be, come from this paper as well.

> it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. (p. 433)

What follows is a description of the Imitation Game, in which an interrogator (labeled C) attempts to decide which gender to assign to two participants (labeled A and B) whom C can communicate with only via teletype (or "chat" in today's parlance). The man is supposed to attempt to convince the interrogator that he is the woman. Turing concludes by saying

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

It is somewhat confusing whether he intended for the machine literally to take the place of a man playing against a woman with the judge deciding which was the woman, or whether he assumed an implicit *mutatis mutandis* clause, so the "man–woman" issue was replaced by "machine–person." I will assume the latter, which is the standard reading.

Notice the way Turing invites us to consider his game. Realizing that it is impossible to define "machine" and "think" in a way that will not beg the question "Can a machine think?," he proposes to "replace" it by a "closely related" question. He never explains what this "close relation" is.

Later in the paper, reviewing progress, he summarizes thus:

> It was suggested tentatively that the question, "Can machines think?" should be replaced by "Are there imaginable digital computers which would do well in the imitation game?" (p. 442)

But "replaced" for how long? To what end? Later he shows his hand considerably more clearly:

> It will simplify matters for the reader if I explain first my own beliefs in the matter. Consider first the more accurate form of the question. 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 that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. (p. 442)

The question requiring definition of "machine" and "think" is, according to Turing, "meaningless." I believe his reason may be found in his earlier remark that answering it would require a "Gallup poll" on what the words meant. By the end of the 20th century, he predicts, when computers can pass the Turing test (a prediction that misfired), "opinion" will have changed so that the outcome of the poll would be different, and one could casually speak of machines thinking as a normal thing. Turing and AI would have achieved an end-run around the question by focusing on winning the Imitation Game as a more tangible research goal.

Does all this show that Turing meant to *define* intelligence (or "thinking") as the ability to pass the Turing Test? It certainly doesn't seem so. The "close relation" one question bears to the other at the beginning of the paper has become rather more distant by the end.

The closest Turing comes to equating "thinking" with passing the Turing Test is in his reply to the "argument from consciousness," which he quotes from Geoffrey Jefferson's Lister Lecture of 1949: "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain..." [7, p. 1110]. Note that the conclusion Jefferson rejects is that "machine equals brain," a conclusion Turing makes clear (p. 451) he does not accept either. The title of the section of Jefferson's paper in which these arguments appear is "Thinking," a word Turing didn't want to use. So it's odd that Turing quotes Jefferson for the argument from consciousness. Perhaps his choice is explained by the fact that, just like Turing, Jefferson is tempted to equate thoughts and words — a temptation for *any* intellectual. Just before the quote Turing extracted, Jefferson says that the "schism" between computing machines

and brains "arises over the use of words and lies above all in the machines' lack of opinions, of creative thinking in verbal concepts" [7, p. 1110]. It sounds to me like the two mens' positions were not that far apart, Jefferson might indeed have abandoned his position if confronted with a machine that could carry on a dialogue such as the one Turing pictures refuting him [22, p. 446].

In any case, Turing's response is clumsy. He actually gives *two* arguments against Jefferson, in rapid succession:

1. "According to the most extreme form of [Jefferson's] view the only way by which one could be sure that machine thinks is to *be* the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course [according to Jefferson] no one would be justified in taking any notice."
2. "Likewise according to this view the only way to know that a *man* thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe "A thinks but B does not" whilst B believes "B thinks but A does not." instead of arguing continually over this point it is usual to have the polite convention that everyone thinks."
   (Turing [22, p. 446], emphasis in original)

The second argument is a rather casual restatement of a standard argument against solipsism: Since it would be impossibly awkward for everyone to walk around acting like a solipsist, let's agree that all people "think," meaning that they are conscious. Turing tries to slip machines in by invoking "the polite convention that *everyone* thinks," but to allow "everyone" to include machines begs the question — the question of consciousness, that is.

The first argument is more novel. It is that, if Jefferson is wrong, it would be intensely frustrating to be a machine and have people, for apriori reasons, refuse to believe that one could be frustrated. Out of simple charity one should give a machine the benefit of the doubt if it could express what it sincerely *believed* to be feelings. Elsewhere in the paper Turing refers to the Muslim "belief" that women have no souls,[3] and says that it seemed to him, as a nonbeliever, that believers must accept that God could give any being a soul if he wanted to. It is not too far-fetched, in the image of the suffering machine, to sense Turing's own frustration at being unable to talk about his homosexuality in terms his society could, psychologically and theologically, accept.

As everyone knows all too well, Turing did not survive his own silent suffering. If he had, we would have had ample time to hear from him on what he really intended by proposing the Turing Test. I think he would be shocked to find a symposium in whose description we read, "[The Turing Test] has become synonymous as 'the benchmark' for Artificial Intelligence in popular culture," but "[d]uring a recent debate at AISB09 . . . , it became clear that, whilst the Turing Test has served well in driving the research agenda in AI forward, it should no longer serve as a true test of intelligence" [1]. There are two reasons to disagree sharply with these statements.

First, the Test has not served particularly well in driving our research agenda forward, and if he were alive today Turing would almost certainly *not* be organizing a team to win the Loebner Prize [19]. Instead, the AI community has worked on a series of much narrower problems, which is disappointing to those hoping for "general artificial intelligence" to crystallize suddenly. In the nineteenth century, visionaries may have fantasized about someday understanding biology well enough to create life, but the only one who actually tried was the fictional Frankenstein. For the real scientists, it was a triumph to synthesize urea, even if this achievement was only a step along a long road, whose twists and turns could not be imagined until it was actually followed.

Second, no AI researcher should accept that the Turing Test has served as "a true test for intelligence." (I picture robots being brought before an examiner who would administer the True Test; those who failed it would perhaps be consigned to the assembly lines. It sounds so medieval, like seeing if witches can float.) Intelligence in a field is the ability to out-think one's competitors in that field; intelligence in a tennis player is the ability to fool their opponent into guessing wrong about where they are sending the ball. A "true test" for this kind of intelligence is a tennis match. I meet the objection that this leaves us unable to test for "general intelligence" by replying that I doubt there is such a thing. If we are interested in conversational intelligence about some topic, then a test might be whether someone who requires information or help about that topic prefers to talk to competitor A rather than competitor B. If we are interested in conversation about nothing in particular, then the question is whether a person would rather talk to competitor A or competitor B. Or a judge might look at a dialogue between A and B and decide which one had made wit-

[3] It is a myth that Muslims believe this, although their actual beliefs about women's souls aren't very palatable either [20].

tier use of their "opponent's" utterances. (Some future program might compete against a future Dorothy Parker or Oscar Wilde.)

In any case, a test of intelligence is always a test of "more intelligent than" rather than simply "just as intelligent as." As Hayes and Ford point out [6, p. 973], putting it the latter way makes success depend on inability to detect a difference: "This is such a fundamental error that it has been given a title: confirming the null hypothesis."

I think Turing would have made clear that he never intended his imitation game as an institution, a *permanent* replacement for substantive questions about thinking. He just wanted to talk about playing this game as a more productive enterprise than talking about "thinking" as things stood *in 1950*. It is tragic that he was foreclosed from clarifying his intentions in later years because he was hounded to death by prosecution, hormone injections, and public humiliation. This tragedy is a burden the AI community has had to bear, one of many little and not-so-little injuries homophobia has inflicted on humanity over the centuries. Unfortunately, it looks like we're going to be bearing it until we either succeed in creating a machine that will realize Turing's vision and make quibbles about whether it thinks seem silly, or admit that it can't be done.

3 Philosophy and Cognitive Science

It is a puzzle why an essentially *philosophical* question about improving a test we don't need should be discussed at a *scientific* conference. It raises the more general question about what the proper role of philosophy should be within cognitive science, or, more generally, what the relation is between philosophy and science. It is a commonplace that sciences split off from philosophy as they *become* sciences. Physics ceased being philosophy in the seventeenth century; psychology, in the twentieth. Chemistry split off from both alchemy and philosophy at the same time, in the late eighteenth.

In Restoration England, there was serious debate about whether getting your hands dirty with experiments was the right way to understand the physical world, versus apriori reasoning within frameworks such as classical geometry [18]. The experimenters won the debate, and ever since philosophers can't sit at the science table unless they have an experiment to propose or a piece of mathematics to contribute.

That's a bit harsh. You could also argue that philosophers have a job to do in clearing up conceptual confusions. For example, philosophers such as Hilary Putnam rightly cried foul at the widespread complacency that the Copenhagen Interpretation brought to the interpretation of quantum mechanics [13] for decades. It might be thought that philosophers' many contributions to the literature on the Turing Test are just part of an effort to help us understand it better.

And yet, when you look at all the papers on the Test (see [11] for a recent sampler), the overwhelming majority them seem either to present scenarios in which some entity passes it but is not intelligent after all, or to contain counterarguments regarding those scenarios. The skeptical papers are inspired by the conviction that whatever underlies human intelligence, it can't be computation. Hence the desire to find a way to prove apriori that computation can't pull the weight it's supposed to. Alas, such apriori arguments work only if you have a definition of intelligence. If intelligence is a natural phenomenon, then it can't be defined, any more than "malaria" can be. Before anyone knew what the natural history of malaria was, the word "malaria" just meant, "this disease with characteristic symptoms whose cause we suspect to be bad air" (hence the name). We can now identify malaria with the disease caused by a certain parasite transmitted through mosquitos. But this isn't a *definition*, any more than water is *defined* to mean "H_2O."

Turing was sophisticated enough to avoid trying to define intelligence, or even give a sufficient condition for it. But others have fallen into this trap, including AI textbook authors such as Ginsberg [5]: "AI is the enterprise of constructing a physical symbol system that can reliably pass the Turing Test" (p. 17). As soon as enough people had misinterpreted Turing, the literature just grew from there.

This has happened more than once. One example is the odd history of the "frame problem." This was originally a technical issue of representing change using predicate calculus [10]. However, it was misunderstood, initially by Dennett [3], as having something to do with collecting all and only relevant information to reason with, and soon, under Fodor's magisterial supervision [4], it had mushroomed into the problem of scientific induction! Other authors define it differently, in a myriad of ways [2, 14], because the philosophers' version is not a real problem, just a good screen on which to project a set of vaguely defined objections to the possibility of a mere algorithm being able to do as good a job as people do at collecting all and only the relevant information to consider in a given circumstance. Anyone who is interested can now look up the answer to the actual frame problem [15, 16], but the philosophers' frame problem has taken on a life of its own. (For an excellent overview of both, see [17].)

Philosophers can, of course, do what they please. The question is why cognitive scientists should pay them special attention. Is there some research program in AI that is fundamentally misconceived for reasons philosophers can explain? Specifically, is there some problem connected to the Turing Test that bears on the current research agenda of AI? Of course not.

4 Conclusions

There is no need to find an improved version of Turing's test for intelligence, because this test as it is conceived of today is largely the creation of philosophers for their own purposes. Turing certainly never intended his thought experiment to provide a synonym for "intelligence" (or "thinking"), or even to provide a practical way of testing for the presence of intelligence. He was writing for an audience in 1950, when the idea of machines that could think seemed preposterous. In today's conditions his thought experiment appears to be redundant — just as he predicted it would be. It's time to lay it to rest.

REFERENCES

[1] Aladdin Ayesh. Call for papers, AISB 2010 symposium on topic: Towards a comprehensive intelligence test (TCIT): Reconsidering the turing test for the 21st century, 2009.

[2] *The Frame Problem in Artificial Intelligence*, ed., Frank R. Brown, Morgan Kaufmann, San Francisco, 1987.

[3] Daniel Dennett, 'Cognitive wheels: the frame problem in artificial intelligence', In Pylyshyn [14].

[4] Jerry Fodor, 'Modules, frames, fridgeons, sleeping dogs, and the music of the spheres', In Pylyshyn [14].

[5] Matthew L. Ginsberg, *Essentials of Artificial Intelligence*, Morgan Kaufmann Publishers Inc, 1993.

[6] Patrick Hayes and Kenneth Ford, 'Turing Test considered harmful', in *Proc. Ijcai*, volume 14, pp. 972–977, (1995). Vol. 1.

[7] Geoffrey Jefferson, 'The mind of mechanical man', *Brit. Med. J*, **1**(4616), 1105–1110, (1949).

[8] Simon Lavington, *A History of Manchester Computers, second edition*, The British Computer Society, Swindon, 1998.

[9] David Leavitt, *The Man Who Knew Too Much: Alan Turing and the Invention of the Computer*, W.W. Norton & Co., Atlas Books, New York, 2006. "Great Discoveries" series.

[10] John McCarthy and Patrick Hayes, 'Some philosophical problems from the standpoint of artificial intelligence', in *Machine Intelligence 4*, eds., Bernard Meltzer and Donald Michie, 463–502, Edinburgh University Press, (1969).

[11] James H. Moor, *The Turing Test: The Elusive Standard of Artificial Intelligence*, Kluwer Academic Publishers, Dordrecht, 2003. Studies in Cognitive Systems.

[12] Henry O'Connell and Michael Fitzgerald, 'Did Alan Turing have Asperger's syndrome?', *Irish J. of Psychological Medicine*, **20**(1), 28–31, (2003).

[13] Hilary Putnam, 'A philosopher looks at quantum mechanics', in *Mathematics, Matter and Method. Philosophical Papers* (2nd edition), ed., Hilary Putnam, Cambridge University Press, (1979).

[14] *The Robot's Dilemma: The Frame Problem and Other Problems of Holism in Artificial Intelligence*, ed., Zenon Pylyshyn, Ablex Publishing Co, Norwood, N.J., 1987.

[15] Raymond Reiter, *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, The MIT Press, Cambridge, Mass., 2001.

[16] Murray Shanahan, *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*, MIT Press, 1997.

[17] Murray Shanahan, 'The frame problem', in *Stanford Encyclopedia of Philosophy*, (2009). the.

[18] Steven Shapin and Simon Schaffer, *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*, Princeton University Press, 1985.

[19] Stuart Shieber, 'Lessons from a restricted Turing Test', *Comm. ACM*, **37**(6), 70–78, (1994).

[20] Jane I. Smith and Yvonne Y. Haddad, 'Women in the Afterlife: The Islamic View as Seen from Qur'an and Tradition', *Journal of the American Academy of Religion*, **43**, 39–50, (1975).

[21] Alan Turing, 'On computable numbers, with an application to the *entscheidungsproblem*', in *Proc. London Math. Society* , volume 2-42, pp. 230–265, (1937).

[22] Alan Turing, 'Computing machinery and intelligence', *Mind*, **49**, 433–460, (1950).

[23] Alan Turing and R.A. Brooker. *Programmers' Handbook (2nd Edition) for the Manchester Electronic Computer Mark II.*, 1952.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Computer system that likes chess

Pawel Dybala[1], Rafal Rzepka[1], and Kenji Araki[1]

Graduate School of Science and Technology
Hokkaido University
{paweldybala, kabura, araki}@media.eng.hokudai.ac.jp

Abstract. In this paper we present our approach to the old computer science dilemma: is proper behaviour enough for a computer system to be considered intelligent or human-like? We discuss the two most classical approaches to this problem, presented by Turing and Searle, and present our own argument by analyzing a simple situation, possible to happen in a real life. We discuss the consequences of taking such approach as ours, present the whole problem from the eyes of an average user, point out some future directions for the field of HCI and propose how Turing Test could be used to achieve some important goals in nowadays science.

1 INTRODUCTION

In recent years we could see numerous projects aiming at creating human-like, conscious machines. A good example of such a venture is the LIREC Project [1], in which European researchers are working on creating artificial companions able to build long-term relationships with humans.

The exciting idea of constructing such machines is what makes people want watch so many science fiction movies. It is also one of major depressive factors for AI scientists, when they discover that it is impossible to construct such machines, as they only operate on zeroes and ones, and even if we make them work also on threes, fours or even hundreds, it will not change the fact that it is hard to imagine mathematic operations conceiving consciousness. Despite this, many researchers, not only from the field of AI or computer science, but also philosophers or cognitivists, still continue endless discussions on how to construct machines that would think or be conscious.

As a matter of fact, this problem is not restricted only to consciousness of computers. In recent world of science we can see numerous research projects aiming at constructing machines that think, talk, experience emotions, are able of liking or disliking things etc. Also here it is quite easy to fall into depression, when you realize what computers really are and how incapable of doing things like these they will always be. Actually, they cannot even perform such simple actions as adding numbers, which was explicitly showed by Levesque [2]. Thus, how can we even think of making machines that can think?

This makes the whole idea of conscious computers seem not worth fighting for. The very basic concept of AI loses its sense, as one may think that constructing truly intelligent machines will never be possible.

This pessimistic argument is not new, as its various mutations were presented by many scientists over decades. One of the best known ones is the Chinese Room thought experiment, in which John Searle (criticizing Alan Turing's concept) claimed that it is possible for to act as if one knew Chinese, which requires only a book with sets of proper rules for this language [3]. This is obviously similar to what computer systems do, making Searle state that behaving AS IF one was doing something is not equal to actually doing it.

Both Turing's and Searle's arguments have been widely discussed, dividing the world of science. Turing's followers claim that getting the behaviour right is enough for the system considered to be intelligent, while Searle's followers argue that acting as if does not equal doing things.

In this paper we would like to present our approach to this subject. Both Turing's and Searle's arguments base on their thought experiments – however, such experiments have one serious drawback, namely: they are theoretical. One of the main sources of criticism towards these approaches is that it is virtually impossible to actually recreate their settings (e.g. the Chinese room) in the real life. Thus, we decided to illustrate our approach with something much simpler – an analysis of a hypothetical, but very trivial situation, likely to occur in the real life. By this, we show how human-computer interaction can be similar to human-human interaction on a very basic level. We discuss conclusions coming from this fact and point out their practical consequences for the world of AI.

2 THE TURING TEST AND THE CHINESE ROOM

In his work published in 1950, Alan Turing [4] stated that getting a computer system's behaviour (in the long run) right and making it indistinguishable from human should be enough for the system to be considered human-like. This is what the famous Turing Test was designed for – to conduct experiments investigating if human judges are able to tell the difference between computers and humans.

Although more than half a century old, this idea was and still is widely discussed. The most influential argument against it, known as the Chinese Room, was presented by John Searle [3]. He proposed a thought experiment, in which a monolingual English speaker is locked in a room and given a book in English, containing a set of all rules needed to process input and generate a proper output in Chinese. The person in the room is given pieces of paper with Chinese characters, and, using the book, processes it to create an adequate response.

This thought experiment, says Searle, proves that one can imitate being able to speak Chinese, without actually knowing it, using only the set of rules. Thus, getting the behaviour right is not enough for anything to be considered conscious.

The Searle's approach, despite having a huge impact in the world of science, causing even questioning the basic concept of AI, was also criticized throughout the 30 years since it was proposed. Levesque, for instance, [2] argued that in the real life the "magical" book with all necessary rules for Chinese cannot exist. He proposed another simple thought experiment, called the Summation Room, in which he argues that the complexity of any set of rules, allowing imitating human behaviour in the longer run makes it virtually impossible to construct machines that would actually do that, due to the enormously long time it would take to process such amounts of information.

French [5] goes even further in criticizing the Chinese room, saying that the question underlying the argument is wrong in the first place. Instead of asking "What are the implications of the fact that someone is answering questions in Chinese without knowing Chinese", we should ask if the very idea of such situation makes sense at all [5]. Needless to say, to French the answer is no, as he argues that the whole setup of the thought experiment is impossible to recreate in the real life, also stating that no such rule-books can exist in the first place.

As we can see, there is still much doubt concerning these two ideas, and the discussion is still vivid. Thus, with this paper we would like to contribute to this field, showing our approach to Turing's and Searle's conceptions.

3 COMPUTER SYSTEM THAT LIKES CHESS

As mentioned above, thought experiments are too theoretical to be fully trusted. Thus, in this section we are going to analyze a very simple, life-like situation, and next draw some conclusions from what we find out.

Imagine a situation in which two people first meet – let us call them Human A and Human B – and all they do is talk to each other. An example exchange of utterances may then look like this:

Human A: So, do you like chess?
Human B: Oh yeah, I do!

Hearing that, Human A will most probably think something like "OK, if Human B says so, he/she probably does like chess", as, looking only at the utterance, there is no particular reason why Human A would think different. Thus, Human A's knowledge is derived on what Human B said, and a logical consequence of this is to say that to Human A, Human B likes chess.

Now, let us imagine a similar situation, in which Human A is talking to a computer system – let us call it System C. If Human A uses the same approach and asks the same question – "Do you like chess?", and the System C answers in the same way Human B did ("Oh yeah, I do!"), logically speaking, Human A should assume that System C does like chess, as it says so.

Thus, in a very simple way, a computer system that likes chess – or is believed to like chess – can be created.

4 DISCUSSION

Needless to say, above experiment would not work like that in the real life – not today, anyway. The main reason why it would not is that humans are aware of the fact that machines cannot like things, as they are machines, not humans.

This is, however, where the Turing Test should prove useful. If the dialogue between Human A and System C was conducted after hiding the latter's identity, say – occurred on an on-line chat channel, Human A would then have no particular reason[1] to doubt System C's words. Thus, in Human A's reality, System C would like chess.

The problem underlying this approach is not new. The knowledge of the fact that the conversation partner is not human was the main issue of the Turing Test concept. In fact, it can be said that in comparison with humans, computer systems start from a much worse position – when we talk with humans, we do not wonder how they are build, how their brain works or what they have inside. Instead, we just assume that they think, feel and like things, as we know that this is how humans are. Contrary to this, in case of computers we do wander how they do what they do, especially if they act as humans. In fact, we could expect that the more human-like computers act, the more suspicious about their abilities humans would be – as they are computers, so, to common knowledge, they should not behave like humans. Even a human behaving in a very odd way would probably be considered more human-like than a perfectly human-like machine.

Thus, it seems like it is all a matter of the right approach. If we manage to convince people not to doubt if creating human-like computers is possible, and to take only their behaviour into consideration, they most probably would start to believe that their artificial partner in fact are human-like, being able to like or know things like we do.

As a matter of fact, on a very simple level it is already happening. Being able to process only zeroes and ones, computers are believed to know how to add, divide and perform much more complex mathematical operations, just judging on the basis of their output. Word processors use thesauri to suggest us what expressions we should use, making an average user think that the software actually understands what they are writing. Average user does not think how the processor or machine is built or how does it perform its operations. All they care about is a proper output, and the fact that it is obtained using methods different from those we use is not even an issue here.

So, here we face the old, good question: is behaviour everything? Turing, in short, claimed that it is, and Searle – the opposite. Both of them supported their theories with strong arguments – which, however, were of purely philosophical nature. We would like to consider the problem looking from the viewpoint of an average user – the same one that does not wonder how the word processor or calculator is built, as long as the output is right.

We say - if we are aiming at constructing artificial companions for humans, i.e. systems that would be able to perform conversations with us, all we should worry is their behaviour. The very term "human-likeness", so often used in

[1] Apart from obviously limited credibility when interacting with someone only through the Internet

HCI and AI in general, means that systems we are trying to construct, should be "like humans", not necessarily identical to humans. If the system acts as if it liked chess, what is the reason an average user should not think that it actually does?

Needless to say, simply saying "I like chess" is not enough. Both Turing and Searle agree that the behaviour should be proper in the long run and on more than one level. Saying "I like chess" may work for the first impression, but if the interaction continues, the utterance itself becomes insufficient. Levesque [2] gives an example of a conversant shouting "I love the Yankees!", which may appear as a manifestation of actual feelings, but is meaningless if nothing interesting follows it [2]. Thus, a system that could be recognized as liking chess would have to act as if it liked chess also in other ways – by, for example, making it topic of conversation from time to time or displaying knowledge about the discipline. If we are talking about embodied agents (robots, humanoids, virtual agents), they could present their fondness also in other ways, for instance - by wearing T-shirts or caps with chess figures. If such sets of behaviours would be recognized by humans as indication of the system liking chess, what is the reason for them not too believe it?

As mentioned above, the main remaining issue then is the problem of knowledge regarding the partner's identity. We know that computers are different from humans, so even if they behave like us, they cannot be the same. As a matter of fact, this starts to become a vicious circle – computers differ from us, so they must behave different, which in turn would make them appear even more different.

Therefore, what we really need here is a break trough in humans' approach, a change in our way of thinking. This could be achieved by misguiding users and making them interact (in the long run and on multiple levels) with non-human partners hiding their identity from them. If the interaction goes well and users would believe that they interact with humans, the true identity of the partner could be revealed. This, hopefully, could convince at least some humans that if computer systems behave in the right way, they could be treated in a similar way to humans.

These, we believe, should be the new challenges for the Turing Test in the 21st century: to find proper methods to convince average users that computers can behave like humans. What we proposed in the above paragraph is only a concept of what it could look like – it still requires empirical confirmation and, when confirmed, more detailed adjustments, like how long should the interactions be and what they should contain. In short – we need to check if this approach actually works, which is possible, contrary to thought experiments like the Chinese room.

If the approach does work, and we actually find a way to make users think that computers can behave like us, it will mean that in our projects aimed at building virtual companions, we can focus solely on their behaviours.

5 CONCLUSION

One may wonder what the above argument is actually about. Making right behaviour the main issue of HCI is not indeed a new idea. Yet, probably due to arguments such as Searle's Chinese Room, in the world of science we sometimes tend to forget it and lose ourselves in endless discussions like "how to make our systems truly intelligent or possessing knowledge". This even becomes the main topic of various symposia and workshops, such as "IJCAI'97 Workshop: Animated Interface Agents: Making Them Intelligent" or "IJCAI'09 Workshop: Knowledge and Reasoning in Practical Dialogue Systems". In discussions conducted during such events some participants tend to present rather pessimistic approach, saying that making computers intelligent or conscious is by all means impossible, as they are nothing more than very fast calculators. Quite popular saying among AI researchers is: "we do not have to worry how we will feel about it when we construct AI, because it is not going to happen".

If we take into consideration the fact that computers are only about zeroes and ones, it can be true – we will never construct AI in the way we imagine it. This approach can make some researchers resign from even trying and abandon their projects – after all, what is the point in pursuing an impossible goal?

However, if we think of computer systems' behaviour as the main issue, we do not have to be such pessimists. If users start to appreciate systems' acting AS IF and start to recognize them as able to be human-like, they will (probably) be satisfied with their new artificial companions. Thus, what we should focus on in the nearest future is:

1) getting the systems behaviour right – so that in all possible dimensions they would resemble humans
2) convincing the users that the fact that their partner is non-human does not have to mean that they cannot behave like us.

If we succeed and manage to construct computers that would act as if they, say, likes chess, which would be recognized by users as such, it would be quite an achievement. And if we do that, who knows – maybe if we create something that behaves in an intelligent way, we will realize that we start to treat it as if it really were? And then, how different for us it would be from interacting with actually intelligent partners?

REFERENCES

[1] LIREC - LIving with Robots and InteractivE Companions, http://www.lirec.org/

[2] H. J. Levesque. Is it Enough to Get the Behaviour Right? In Proceedings of T21st Joint Conference on Artificial Intelligence (IJCAI-09), 1439-1444, Pasadena, California, USA (2009)

[3] J. Searle, Minds, brains, and programs. *Brain and Behavioral Sciences* 3, 417-457, 1980.

[4] A. Turing, Computing machinery and intelligence. *Mind* 59, 433–460, 1950

[5] R. French, The Chinese Room: Just Say "No!", *Proc. of the 22nd Cog. Sci. Conf.*, 657–662, Philadelphia, 2000.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Man, Machine, and Interpretation.
Donald Davidson on Turing's Test

Stefan Riegelnik
Stefan.Riegelnik@wu.ac.at

In the year 1950 Alan Turing published his influential paper "Computing Machinery and Intelligence", where he presented at first the idea of a machine, which is able *"to carry out any operations which could be done by a human [being]"* and insofar "*can in fact mimic the actions"* of a human being *"very closely"* (Turing 1950). According to Donald Davidson the remarkable feature of this idea is that it presupposes what a lot of people in the following discussion have missed as a decisive aspect of the behaviour of the machine: namely to understand what it does, i.e. to have an understanding of the meaning of the signs which the machine manipulates. This assessment is in some way similar to a lot of doubts about the fruitfulness of Turing's test by philosophers concerned with language and mind. In contrast to John Searle, as one of the prominent critics of Turing's idea, for Davidson this does not show the deficiency of the idea of the machine as a criterion for having thoughts. It rather proves that questions concerning the conditions of having thoughts depend on questions concerning the conditions of attributing thoughts. So when Donald Davidson explicitly focused on Turing's Test (1990) as a thought experiment his assessment was not at all completely negative, but rather ambiguous: on the one hand he regards the test as inadequate, on the other hand he concedes that it reveals some important aspects concerning the attribution of thought. My contribution to the still ongoing discussion about Turing's Test is meant to emphasize the significance of the argumentation by Donald Davidson as a philosopher who has been regrettably neglected in the assessment of Turing's test.

*

The philosophical relevance of Turing's test does not seem apparent at first sight, since one could argue that nothing follows from asking people, i.e. asking whether they believe that some thing thinks: if on the contrary one believes it does, then the fact that an object "succeeds" in performing in a way, that is assessed by the interrogator as a way human beings act and talk, thereby scoring a certain amount of points, then one could rightly draw the conclusion that the object is what other people take it to be meaning in this context that the object *thinks*, because human beings think that it thinks. This is especially plausible since "alternatives", like a "direct" way to another person's thought is not available. Thus one could point out that Turing's test serves indeed as a proof that a machine can mimic the behaviour of a human being, showing thereby that Turing's test is an adequate substitution for his initial question "Can machines think?".

The focus on this "third person point of view" is one reason for which Turing's test has been criticized. Davidson stresses the question of the presupposed truth of behaviourism and points out, that since "*the Test is to see whether, and on what basis, normal human judges are able or willing to assign mental attributes to machines*", behaviourism "*in the broadest sense must be assumed*". According to Davidson, this cannot be objected to Turing's Test, since "*the same question will apply equally to everyone (else)*" (Davidson 1990): no one has access to the thoughts of a thinking being without observing the behaviour. Talking about thought requires adopting a third person point of view. Davidson is strictly against being able to settle the question whether an object can think *without* being able to say anything about the *content* of her/his thoughts. I think this is important to mention here because in the history of philosophy many attempts have been made to explain the nature of thought that ignore the fact that thoughts without content are "empty". A view opposed to this ends always up with comparing thoughts with rivers, streams (anything in a flux), or with a theatre. What "unifies" these metaphors is the fact that essential features of thoughts cannot be explained, or even worse, as Mras shows in her work on Nietzsche and Wittgenstein, these metaphors are rather misleading than helpful. The possibility of ignoring the content is excluded in the scenario of Turing's test – "*any evidence that thinking is going on will have to be evidence that particular thoughts are present*" (Davidson 1990). Davidson deems this as a particularity of Turing's test – he contrasts this with normal circumstances where it is not necessary to discover *what* a human thinks to know *that* it thinks "*by being told, or knowing in some other way that it is a man or woman*" (Davidson 1990). Although Davidson weakens this by holding that these "*are fallible ways of knowing*", this line of argument makes Davidson's position vulnerable to arguments that Turing's test is a test for *human* intelligence only.

*

In order to show that even from a third person point of view a sceptical attitude towards the equation of human and artificial intelligence is justified, I am going to examine the conditions of ascribing meaning to purely mechanically produced signs. Considering the fact, that Turing's test is an artificial situation, Davidson admits that many aspects of human behaviour have to be removed from consideration leaving solely the occurrence of signs as expressions of what humans do. In other words, the occurrence of the signs alone could enable the interrogator to decide whether the sign-producing object has thoughts or not, which comes down that "*[t]he Test makes meaningful verbal responses the essential mark of thought*" (Davidson 1990).

The focus of Davidson's argument lies at first on the interdependency of "having thoughts" and "the signs" or better, "the meaning of the signs". He writes, "*if a response has any linguistic meaning at all, there must be a thoughtful cause. But the computer is*

the thoughtful cause of a response only if it means something by the words it produces" (Davidson 1990). Only if the interrogator can identify the signs appearing on the screen as linguistic utterances *and* if he is able to find out what they mean, he knows the content of the thoughts expressed and consequently, *that* the machine thinks. This line of reasoning reminds one of Searle's thought experiment known as the "Chinese Room Argument". For Searle the fact that signs occur on the screen does not show that the machine that "does this process" exhibits thereby a meaningful behaviour. What computers according to Searle do is at most a formal manipulation of signs. Davidson argues in a similar way: *"[…] Turing has his computer produce what seem to be English sentences. But of course the computer might not be speaking English. Those sentences might have entirely different meanings for the computer than they have in English; or they might have no meaning at all*" (Davidson 1990).

The question of whether there is a relation between thought and talk and whether the relationship is necessary are here of minor interest; the pivotal question is now whether the interrogator can find out the meanings of the "signs" produced by the machine. This is essential, since the signs are the sole evidence to find out *what* the object thinks and consequently, *that* it thinks.

*

One might argue here that too much is taken for granted, since how can we know that the results of machine's operations are to be understood as *signs* at all? The only way to settle this question is to take up the effort to understand the expressions of the machine, i.e., to find out the meaning if what might be taken as signs. At first it seems that there cannot be any difficulty involved in this task, since the interrogator could interpret the signs appearing on the screen just as every other utterance. A common objection to this line of reasoning is that a human being has programmed, e.g. an ATM, so the signs appearing on the screen are purely "formal symbol manipulations": the signs are not produced by a thoughtful machine but by a programmer. Davidson writes: *"[…] it is the programmer who has supplied the semantics, who has understood English, and has given meaning to the words produced by the object*" (Davidson 1990). Nonetheless one could still ask: how can we exclude that it is the machine that caused the utterances occurring on the screen? The peculiarity of this question is that it cannot find an answer without presupposing the question already having been answered.

For this reason, Davidson modifies Turing's Test. The new scenario has a single object under scrutiny. The interrogator now has to find out whether the object thinks or not by trying to interpret the signs produced. This test is different and so are the requirements in order to satisfy it, but bearing in mind the fact that Turing's Test eliminates the possibility of telling whether an object thinks without knowing what it thinks, Davidson's Test is as

suitable as Turing's, with a minor adjustment: since Davidson understands Turing's test as a way to explore the nature of thought, the modified test is more suitable "*if what we want to know is what our (the interrogator's) criteria for thought are*" (Davidson 1990). For reasons of alienation, it would be useful to have the machine uttered words in an language unknown to the interrogator, but as Davidson holds elsewhere, *"[t]he problem of interpretation is domestic as well as foreign: it surfaces for speakers of the same language in the form of the question, how can it be determined that the language is the same?*" (Davidson 1973).

If we assume that the signs on the screen are English sentences (Davidson cites a correct syntax and correct relations between questions and answers in favour of this assumption and argues that "*we have no better evidence if the object were an English-speaking person*"), the question is whether the interrogator is able to interpret these signs. Davidson denies this idea because the interrogator could not have a *"clue to the semantics of the object*". The interrogator cannot tell, in spite of the correct syntax and the harmonization of questions and answers, what the signs mean since "*[t]here is no way he [the interrogator] can determine the connection between the words that appear on the object's screen and events and things in the world*" (Davidson 1990). But why not? Davidson mentions two methods of finding out what people mean by what they say: the "indirect" method, whereby one identifies a person who learned a language "*through the usual conditional process, which connect things and events with words*" (Davidson 1990), and the "direct" method, which is the sole available in the test scenario, where "*we can discover these connections directly, through observing relevant causal interactions between the speaker, the world, and the speaker's audience*" (Davidson 1990). The latter is very much the same as Davidson's thought experiment "radical interpretation", which is intended to pinpoint the required knowledge for being able to interpret linguistic utterances. The radical interpreter can rely on no prior knowledge of the speaker's beliefs or of the meanings of his utterances: he has to interpret the utterances *from scratch*. The only material available to the interpreter is the speaker's behaviour towards objects and events. To describe the process of radical interpretation and, simultaneously, the attribution of thoughts, Davidson suggests to understand the attribution of thoughts in the model of a triangular structure: "*If the two people […] note each other's reactions (in the case of language, verbal reactions), each can correlate these observed reactions with his or her stimuli from the world. […] The triangle which gives content to thought and speech is complete.*" (Davidson 1991).

*

Davidson thus regards Turing's Test as inadequate not because "*the Test restricts the available evidence to what can be observed from outside"* but *"because it does not allow*

enough of what is outside to be observed" (Davidson 1990). The interrogator is not able to find out what the signs on the screen mean, since "*[w]hat is needed is evidence that the object uses its words to refer to things in the world, that its predicates are true of things in the world, that it knows the truth conditions of its sentences*" (Davidson 1990). In consequence, Davidson suggests a modification of Turing's Test: "*[t]he object must be brought into the open so that its causal connections with the rest of the world as well as with the interrogator can be observed by the interrogator*" (Davidson 1990). However, it is an open question whether a suitable test scenario along these lines could be designed at all. So either the "*object brought into the open*" must resemble a person in all important aspects, or if it does not, it is clear from the start that the machine "described as mimicking the actions of a human being very closely" does so because it is meant to deceive the interrogator.

References

Davidson, D. (1973): Radical Interpretation. In: *Dialectica*, 27, 212-328.

Davidson, D. (1990): Turing's Test. In: Newton-Smith, W.H. / Wilkes, K.V. (eds): *Modelling the Mind*. Oxford University Press. 1-11.

Davidson, D. (1991): Three Varieties of Knowledge. In: Phillips Griffiths, A. (ed): *A.J. Ayer Memorial Essays: Royal Institute of Philosophy Supplement*. Cambridge University Press.

Mras, G. (2010): Nietzsche and Wittgenstein. Taking on different perspectives. In: Ramharter, E. (ed): *Influences on Wittgenstein*. (forthcoming 2010).

Searle, J.R.: *Intentionality. An essay in the philosophy of mind*. Cambridge: Cambridge University Press. 1983.

Turing, A.M. (1950): Computing machinery and intelligence. In: *Mind,* 49, 433-460.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Panel Communication

Some Misconceptions Regarding the Turing Test

Hugh Loebner

Abstract. Several misconceptions are presented regarding Alan Turing's "Imitation Game" These relate to (a) using other than three entities, (b) using an incorrect criterion for passing the test and (c) unnecessarily restricting the content of the test.

1 INTRODUCTION

I have directed or otherwise been associated with 19 annual Loebner Prize Competitions. During this time I have been the recipient of a multitude of communications regarding the Turing Test. Most of these excoriated the competition, detailing in what manners the Loebner Prize fails as a Turing Test. My association with the Loebner Prize and my reflections on these communications lead me to recognize that there are several misconceptions regarding the test, or, at least, a failure to appreciate salient characteristics.

The Turing Test is a paired comparison between a human and a computer program judged by a human. The Method of Paired Comparisons is a well studied research tool that permits the development of an interval scale and can test the transitivity of an underlying dimension.

2 IMPROPER NUMBER OF PARTICANTS

Turing wrote "It [the imitation game] is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex" [1] and "We now ask the question, "What will happen when a machine takes the part of A in this game?" Note that the judge is presented with precisely two entities which are to be compared against each other on how "human" they seem. In order to conduct a proper Turing Test one must be able to count to three. One must gather a judge, a human, a computer and then Halt. One could say that this is another type of "Halting Problem."

- One error occurs when the creator of a chatbot claims that his or her entry "has passed the Turing Test" if someone interacting with an unpaired chatbot mistakes it for a human. This is not passing the Turing Test because the chatbot was not chosen as the human when compared with a human. I would say this is the most frequent error made by the lay community.

Consider also the "CAPTCHA" device used by many websites to insure that a human and not a web bot is sending an email or otherwise using the site. CAPTCHA is an acronym for "Completely Automated Public Turing Test To Tell Computers and Humans Apart"[1] However, it is not a Turing Test at all.

- A more subtle error happens when a judge is required to judge more than two entities without being told how many are human and how many are computers. This, in fact, was the design used in the Loebner Prize until 2004, and the contest was directed by some very eminent scholars during that period. The initial design of the Loebner Prize was under the direction of a distinguished committee headed by Daniel Dennett, and including, among others W.O. Quine, Joseph Weizenbaum, and I. Bernard Cohen. Nevertheless, the first contest, held in 1991 at The Computer Museum in Boston, MA, did not conform to Turing's prescription of paired comparisons.

Consider Stuart Shieber's critique of the 1991 Loebner Prize Contest [2]. Despite mentioning that the 1991 Competition did not present binary choices, Shieber's complaints revolved around limiting the "tenor" of the comments and limiting "topics" of content. (His main complaint, however, was that the competition was premature.)

When I (along with several others) judged the Loebner Prize in 2001 at The Science Museum in London, the same general design was used. At the time, even I did not realize that I was not participating in a Turing Test. In point of fact, the 2004 competition was the first Loebner Prize wherein the judges were actually faced with a paired comparison. It was only *after* the 2004 competition, while I was reflecting upon it, that it occurred to me that this was actually Turing's design. To be honest, I specified the paired comparison design not because it was the correct design for a Turing Test, but to facilitate holding the competition. In 2004 I held the competition in my apartment in New York City, in part to support my lawsuit against the Cambridge Center for Behavioral Studies[2]. Rather than develop the necessary communications protocols, I required each contestant to enter his own programs to allow interactions between judge, human, and computer[3]. Rule 5 stated:

(5) Each Computer Entry must be accompanied by a matching communications program.. The Com Program must (a) run as two programs on two

[1] http://www.captcha.net/

[2] Loebner v. Cambridge Center, et al., (03 CV 10959 RCL Fed. Dist. Ct., Mass.)

[3] I thought: "Why should I have to do the work of developing a communication protocol?" It turns out the answer is: to prevent contestants from entering a communication program that garbles the human-judge interaction, which actually occurred in the 2004 competition.

> computers (b) communicate character by character with each other via easily implemented communications (c) mimic the appearance and behavior of the matching Computer Entity. "Easily implemented communications" means directly two computers with a cable having standard plugs or connecting each to an Ethernet hub.

I do not claim that requiring one or more judges to select one or more computer programs from among more than two unknown entities is not a test of "intelligence." I am merely stating that it is not the Turing Test. Nor am I claiming that academicians do not know what is the Turing Test. I am just pointing out that (1) the general public does not understand the test and (2) that academicians, at least those who have been associated with the Loebner Prize, seem not to appreciate the central fact that the Turing Test is a Paired Comparison.

2 PASSING THE TEST

I believe that there are two other errors regarding the interpretation of the Turing Test, although these are open to different interpretations of the design.

One error involves what constitutes "passing" the test. Turing wrote "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 that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning."[4] This is sometimes interpreted to mean that mistaking a computer program for a human for five minutes constitutes passing the test. However, Turing did not write words to the effect "...and that would mean passing the test, or, explicitly, "I believe that in 50 years it will be possible to have a computer program pass the test." In my opinion, Turing was simply making a prediction about what would be the state of the art in 50 years, although he did continue ".... I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted," which statement obscures the point. My personal belief is that Turing would have said that a true "thinking machine" should be able to fool the judge for any given length of time.

3 CONTENTS OF THE TEST

Another misinterpretation, I believe, is the view that the Turing Test should be restricted to textual material only. If we adhere to this limitation, we are faced with the limitation that the Turing Test can only be played with a literate person. Illiterates need not apply.

I received a great deal of criticism for requiring that the "Grand Prize" winning program be able to deal with audio/visual input. However, in his article, Turing discusses the memory requirement for a successful program. He wrote "I should be surprised if more than 10^9 was required for satisfactory playing of the imitation game, at any rate against a blind man." Now, we may ask, why should the storage capacity of the computer depend on whether it were played with a blind or sighted person? If we permit visual images when played with the sighted person the answer becomes obvious. And, in the penultimate paragraph of his article, Turing writes:

" We may hope that machines will eventually compete with men in all purely intellectual fields....It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc."

[4] Turing uses 70% chance of success. In a paired comparison, we would expect that a "thinking" computer capable of perfectly imitating a human would only have a 50% chance of success. Also, there is some ambiguity about the five minutes of questioning. Are the five minutes to be divided between the two entities for a five minute test, or is each entity to be questioned for five minutes for a total test time of 10 minutes?

REFERENCES

[1] Alan Turing. Computing machinery and intelligence Mind, 59:433-490, (1950)

[2] Stuart M. Shieber, Lessons from a Restricted Turing Test, *Communications of the Association for Computing Machinery*, volume 37, number 6, pages 70-78, 1994, also available at http://www.eecs.harvard.edu/shieber/Biblio/Papers/loebner-rev-html/loebner-rev-html.html

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

The original test: it's harder than you might think

Darren Abramson

January 11, 2010

Abstract

In this paper I argue against the view that judges can easily distinguish between machines and humans in the Turing test. I show that a particular criticism of this form, as yet unrefuted in the literature on the test, is unsuccessful. In doing so, the paper presents a counterexample to the criticism of the test. I then identify open questions in the philosophy of computer science and demonstrate that the possibility of constructing the presented counterexample to the criticism of the test is reducible to these open questions.

1 Subcognition and the Turing test

This part of the paper involves a short demonstration, requiring the cooperation of able readers.

> Please bring your hands together, palsm pressed together, as if you were praying, touching the fingertips of your left hand with the corresponding fingertips of your of your right hand. Fold down your two middle fingers - and only your two middle fingers - so that the middle knuckles of both come together. (The tips of your thumbs, index, ring and pinky [little] fingers should be still be together.) Now, move your other fingers one at a time and report what happens. (from French (2000))

Can you move your ring fingers? Did you know you wouldn't be able to? Finally, what would happen if you asked this question in the Turing test of the two subjects being interviewed? Before we get to his argument, let us notice what French's demonstration does for our intuitions. Surely NO programmer could ever anticipate every such question, leaving us a toolbox for easily distinguishing human from machine.

In an earlier writing (French, 1990), Robert French imagines applying myriad results from cognitive psychology to the Turing test, by asking questions of the candidates concerning word similarity and semantic appropriateness of novel word strings. French concludes that the physical experience of human beings radically affects cognitive processing in ways that are revealed by these so-called 'subcognitive' questions. I have no argument with *this* claim. However, he goes further to make claims about the necessary relationship between the ability to answer these questions and having had certain experiences.

French suggests asking the two subjects in the Turing test even more salient questions, which will reveal not just cognitive differences which *depend* on physical differences, but will *directly* reveal physical differences (French, 2000) . For example, the subjects can be asked to observe novel properties of their own physiology, which they may not even be aware of until being asked, as we just saw. French anticipates objections that either these performative questions, or subcognitive questions, are not fair game in the Turing test scenario. I have no such objections.

French's claim, in short, is

> . . . as a test of general intelligence, the Turing test is not particularly appropriate precisely because it so hard: it tests not for intelligence, in general, but, rather, for culturally oriented human intelligence. In order to pass it, a machine would have to experience the world in essentially the same manner as we humans had, and, in order to do this, it would have to have a body and a set of experiences very similar to our own. And it is this that would make it virtually impossible for any machine to actually pass the Turing test. (French, 2000, 339).

Rather than rehearse the various subcognitive and embodiment-related questions which French mentions, I offer the following well constrained example which is both illustrative of his argument, and makes clear the issue in question.

> Consider, for example, the question: 'does freshly baked bread smell nicer than a freshly mowed lawn?' A machine that had never smelled either baking bread or a newly mowed lawn would have a great deal of trouble answering this question, unless, of course, it had specifically programmed to answer that particular question. But there are infinitely many such questions that humans can answer immediately because they can make a judgement, based on actual physical experience, about the degree of pleasure associated with each. Further, on average, most people within a particular culture will respond to this type of question in a similar manner. It is this fact that will be used to infallibly trip up the computer. (French, 2000, 334-335).

It is worth noticing that French's characterization of the only reasonable method for programming machines that pass the Turing test is explicitly rejected by Turing. I have shown elsewhere (citation suppressed) that Turing's presentation of the construction of 'learning machines' is motivated precisely by his rejection that computers can only do what we explicitly tell them to – witness the lengthy discussions of Lady Lovelace's objection, for instance. As we will see, the philosophy of computer science has already done a great deal to make issues about what implemented computers can do quite precise.

French elsewhere emphasizes the modal nature of his claim that in order to be prepared to answer certain questions, minds must physically interact in certain ways with the environment. He claims that being prepared to answer questions at the subcognitive level "[is] the product of a lifetime of interaction with the world which *necessarily involves* human sense organs, their location on the body, their sensitivity to various stimuli, etc" (French (1990, 62); emphasis in original).

French's argument is as follows. Necessarily, if you can appropriately answer subcognitive questions then you have had a lifetime of physical interaction with the world. Computers haven't had this lifetime of interaction. Therefore, by modus tollens, computers can't appropriately answer subcognitive questions, and are therefore unable to pass the Turing test. Now, if French's argument is sound, then Descartes has a new tool for allaying his hyperbolic doubt at the end of the first Meditation. Consider the first premise: necessarily, if you can, for example, describe preferences for some smells as opposed to others, then you actually experienced physical interaction with the environment. Forget about hyperbolic doubt – Descartes merely needs to ask himself a few questions about word association, favorite smells, and hand gestures, and he knows there is no evil demon deceiving him about the nature of his existence!

French's argument fails. There surely are possible worlds which contain only these two beings: Descartes and an evil demon. Descartes lives his life as a brain in a vat, and the evil demon makes Descartes think he has physical interaction with the world, when in fact he does not. This vat-Descartes could, presumably, pass a Turing test (or at least a culturally appropriate one) which included all sorts of subcognitive questions.

Someone sympathetic to the subcognitive objection to the Turing test will complain that I have only shown that a metaphysically possible brain can pass the Turing test – this is far from proving that a machine that had not experienced the world as we do could similarly pass it. I will now give a recipe for creating machines which have not experienced the world as we have, yet are able to answer subcognitive questions, and pass the Turing test.

2 How To Build a Machine That Passes the Turing test

1. Pick a person. Draw an imaginary line around their brain and nervous system.
2. Code up a complete model of everything inside the line. That is, for completely deterministic portions of the tissue involved, construct a model which behaves exactly as the tissue does. For non-deterministic portions, have the model behave in identically probabilistic fashion.
3. Code up a model of the effect the following stimuli have on the nervous system just modelled:
   (a) being invited to participate in a Turing test;
   (b) arriving at the test site; (If the model outputs some other decision than participating, erase the session, go back to the first model, and keep simulating until a decision is made to participate. If the model turns out to be recalcitrant, pick someone more cooperative to model.)
   (c) sitting down at the terminal, and seeing questions flashed on a screen in a room just like the one in which the real human subject is sitting.
4. Apply the above stimuli to the model. Copy the model to the computer in the Turing test scenario.

5. For each question, deliver inputs to the brain model in the form of simulated inputs to the nervous system; track simulated efferent outputs to motor neurons, thereby tracking simulated key presses, and send the output to the questioner.

Notice that the computer program described will not only pass the Turing test, answering any variety of subcognitive questions appropriately, but in the case that the program is competing against the 'source' for the model, it *will be indistinguishable from the competing real human being*. The reason for this should be obvious. We have imagined coding a program which faithfully reproduces the outputs of a human nervous system. Then, we have imagined stimulating this simulated nervous system with the inputs a nervous system would receive in taking the Turing test. The only source of difference we might find would be in errors in either simulating the input, or the nervous system itself, and we are supposing no such errors have been made.

Think about the task I asked you to do a few minutes ago. If we can and do write the program I have described, it will respond in just the same we do – with a message something like 'wow, I can't move my ring fingers. I didn't expect that'. This doesn't threaten our intuitions, though. By simulating the muscle contractions, bone movements, and cartilage tension of arms, and sending appropriate inputs to our simulated brain, we received just the same response we would from a real brain connected to real muscles, bones, and cartilage.

Of course, it is an open question whether or not a machine can ever be built, as described, that passes the Turing test. In the remainder of this paper, I will show how we might formulate this question more clearly, and the way in which central issues in the philosophy of computer science bear on it.

3 Really Hard Questions For Machines

It should be emphasized that the program described above can not only pass the Turing test when asked subcognitive questions, it can even – if the programming is done well – pass questions that are so hard, French excuses them from fair game during the Turing test. In his 1990 article, French supposes that the judge prepares him or herself by doing a series of experiments on the psychological phenomenon of 'associative priming' (explained below).

The judge has available a series of macros for sending strings at a preset pace over the terminal hookups in the test, and a piece of software scanning responses from subjects on the screen and measuring latency. Having practiced earlier with a host of human subjects in similar terminal hookups, the judge can come in with statistically reliable predictions on variable latencies for responding to key presses for, say, identifying words from non-words.

It turns out that people respond faster in identifying a string of letters as a word if it is preceded by a semantically related word. As French tells us, "[if] ... the item 'butter' is preceded by the word 'bread', it would take significantly less time to recognize that 'butter' was a word than had an unassociate word like 'dog' or a nonsense word preceded it" (French (1990), 57). The familiar conclusion, that no computer could pass the Turing test, is presumed to follow from the ability of the judge to ask such questions.

> The machine would invariably fail this type of test because there is no a priori way of determining associative strengths (i.e., a measure of how easy it is for once

> concept to activate another) between *all* possible concepts. Virtually the only way a machine could determine, even on average, all of the associative strengths is to have experienced the world as the human candidate and the interviewees had. (French (1990), 58)

Precisely what is left out by the word 'virtually' here? Perhaps French means that, barring exceptional luck and accident, there is no way to build a machine that one expects to pass such tests. He supposes that a critic might object to these questions, since they don't really seem in the spirit of the test. After all, these are explicitly subcognitive questions, and the Turing test is intended to make cognitive distinctions. So, French "obligingly disallows" (ibid.) such questions. I will argue now that no such obligation is necessary. Machines designed as I have described can, if certain programming choices are made, pass these psychometric tests also. Also, by seeing that this is true, we can help avoid a mistake that comes up from time to time in discussions of cognition and computation.

Cummins et al. discuss ways in which you might decide between connectionist and language of thought implementations of cognitive architecture in people. They distinguish between 'primary' and 'incidental' effects of algorithms (Cummins et al., 2001) . Primary effects refer to the computational task a machine or organism accomplishes, while incidental effects are all those facts which may be true or false without affecting whether the task is performed. If one machine emits a beep while factoring natural numbers, but the other emits a squawk, then they might have the same primary effects (factoring numbers correctly) but different incidental effects.

There might be connectionist and rule-based parsers that recognize precisely the same strings, but they argue that there is reason to think that connectionist parsers would be revealed by little or absent growth in processing time compared to input size. However, rule-based parsers, they say, would have some positive correlation between input and processing time.

Here is their summary of the possible sources of difference between two systems in their incidental effects.

> One important source of incidental effects thus lies in the algorithm producing the primary effects. But systems that implement the *same* algorithm may exhibit different incidental effects. If the underlying hardware is sufficiently different, they may operate at greatly different speeds, as is familiar to anyone who has run the same program on an Intel 486/50 MHz [computer] and on a Pentium III [computer] running at 750 MHz. A significant difference in response times between two systems, in other words, does not entail that they implement different algorithms; it may be a direct result of a difference at the implementation level. Hence, two systems that have identical primary effects may exhibit different incidental effects either because they compute different algorithms or because the identical algorithms they compute are implemented in different hardware. (Cummins et al. (2001), 174).

Notice that these authors are not saying that if architecture and/or algorithm are different then there will necessarily be a difference in incidental effects. In fact, they go on to say

that one cannot in general infer from sameness of incidental effects to sameness of functional architecture, which presumable includes both algorithm and implementation.

However, they also appeal to our idea that if you have two completely different functional systems, surely timing will generally go askew between them. Let's look briefly at a counterexample.

I will stipulate the following definition of emulation: one computer emulates another just in case they display both the same primary effects *and* all the same secondary effects, on the definitions offered by above. Furthermore, I will limit secondary effects to those that can be detected by a judge in the context of the Turing test. So, latency between stimuli and responses will count as secondary effects, but the warmth of the processor of the computer being tested will not.

There was a time when video games from one PC platform were unplayable on another. However, in dealing with rapidly increasing processor speeds, programmers tied event timings to some absolute measure instead of clock ticks. With a computer equivalent to a universal Turing machine, the ability to simulate exactly the *timing* of another computational system depends only on the speed at which it operates and the number of instructions it can execute per cycle. Routinely, popular web browsers can display interactive windows in which entire computer architectures in production less than a decade ago are emulated in the sense defined here.[1]

In the next section I will consider objections to the claim that any computer can be built, for the purpose of passing the Turing test, in the manner I have described.

4 Objections

4.1 Brains and their environment cannot be separated, even in principle.

Recent opinions in the foundations of cognitive science suggest such a statement.[2] Consider the idea that human minds are intrinsically embodied and embedded. A strength of French's argument, one might claim, is that it takes advantage of this property of persons. If this objection is taken to mean something like 'if you isolate a brain from body and environment then you no longer have a person', I am likely to agree. However, we are not trying to reproduce people here – we are merely trying to simulate part of a body.

If the objection is taken to mean something stronger, like 'one cannot in principle isolate and emulate a biological subsystem, namely the brain and spinal cord' then objection appears weak. No doubt isolating and emulating any biological subsystem is prohibitively difficult with present technology. For all I know it is impossible – I will talk in great detail below about particular reasons doing this might be impossible below. However, despite continuous interaction between neurons and the rest of the world, it seems at least plausible that future scientists could arbitrarily pick a physical boundary – say synaptic junctions at distal nerve endings – and model everything inside.

4.2 We cannot, in principle, model input to the nervous system.

There are a few problems involved once we have our nervous system modelled. We must be able to specify a physical scenario in somewhat large scale terms, and then translate that scenario into simulated effects on individual synapses. Again, this is a difficult task. A few examples of imaginary beings capable of doing this come to mind: Descartes's evil demon, and the managers of the fictional 'Matrix' of the popular movie by that name. Let us consider this objection in isolation from the one above. Suppose that we are only worried about emulating inputs to a nervous system, whether the nervous system itself is real or emulated. Someone who objects to the possibility of accurately emulating the input that particular environments provide to the nervous system is committed to the success of a judge in what I will call the *Wachowski test*.

The Wachowski test is a lot like the Turing test. However, instead of trying to figure out which chat window is connected to a machine and which is connected to a person, the goal of the judge is to distinguish a person experiencing an *emulated* terminal, entirely computer generated and delivered directly to the nervous system, from the person who really is sitting at terminal typing responses. Again, for all I know, no one, in tandem with the best engineering possible, could produce a judge-fooling entry for the Wachowski Test. Two observations should suffice for rejecting this objection against my counterexample to French's argument, however. First, it is at least possible that no judge can beat the Wachowski test for sufficiently advanced emulations. Second, and now I am considering both objections presented so far, French never mentions *any* reason to think we can't simulate either the environment or the nervous system.

4.3 The Identity Objection

I find this objection the most interesting. It avoids appealing to impossibilities which need to be established empirically, and are far beyond any reasoning we can make from present physics and neuroscience.

Suppose we have chosen to simulate a particular person – say Robert French himself. We conduct the Turing test between Prof. French at one terminal, and the simulation plugged into the environment emulator program at the other end. At the second terminal, text inputs and outputs are received and produced by a program which translates to and from emulated stimuli and responses of the emulated brain.

Recall French's conclusion: to be able to pass the Turing test requires a lifetime of experience involving the world. An objector might claim that the computer in this case, by running an emulation of Robert French's brain *does* have a lifetime of experience involving the world.

While appealing, I argue that this objection – identifying the machine's experience with the experience of its modeling source – is illicit. This can be seen in a few ways.

First, if French's argument is intended to have a general conclusion, then it must deal with computers randomly selected from the space of all possible computers. It is true that given the space of possible computers complex enough to emulate a human brain, very few will produce Turing test-passing output. In fact, very few will produce recognizable outputs at all! However, there will be some which emulate my brain, others that emulate your brain,

and so on for every other brain (that is, supposing other considerations discussed below are decided in a particular way). This is true, even granting that which computers simulate which brains will depend on the program we use for supplying simulated environments. So, while French's brain may play an epistemic role in programming a particular machine, the program itself hasn't had any experience – it waits quietly in Plato's heaven until we retrieve it using a neural scanner, or some other such putative device.[3]

Second, there is a strong sense in which the computer hasn't had any of the experiences which French has had. In a naive sense, the computer – if it had experiences at all – merely had those of being manafactured, plugged in, and then programmed. Its program doesn't even contain mention of interactions with the environment; only numerical solutions to differential equations, perhaps, modelling spike trains given certain stimuli.

Third, experiences presume an experiencer. If we are able to program a computer as I have described, does that imply that computational functionalism, as a theory of qualitative experience, is true? It doesn't seem so. Notice that if we are committed to the computer's sharing French's experiences, then we will be forced to discount the counterexample I have offered at the expense of presuming the truth of computational functionalism. Surely there are possible worlds where computers can emulate brains, and minds – which are the things which do the experiencing – exist as epiphenomenal immaterial souls.[4]

5 Reasons to Think That Brains (and/or environments) Can't Be Emulated

There are two broad sets of reasons which would render my counterexample to French's arguments a failure. First, there could be contingent limits on the computers we can build which would prevent us from every accurately modelling a brain. For example, we might not be smart enough to ever decode the complicated functions being computed by neurons. Or, a cataclysmic solar flare might incinerate the world before we finish working on our best models. These considerations are beside the point, though. French's argument is about the in-principle limits available for constructing machines that pass the Turing test.

More interesting are the possible failures of two related theses. The Church-Turing thesis says that all of the effectively computable functions are Turing computable. This means that all of the functions which a human being can compute by using unbounded time, paper, and pencil, and simple rules about writing, recognizing, and changing symbols, can be computed by a regular Turing-equivalent computer. We can consider an expansion of this thesis, and call it the Physical Church-Turing thesis (as, for example, Deutsch (1985) does). This says that all physically computable functions are Turing computable. Now, suppose that the thesis fails, and that human nervous systems compute functions which no Turing-equivalent computer can. Then no matter how much time, ingenuity, and material support are available, no counterexamples of the form I have presented are possible.

The second thesis is an extension of the Church-Turing thesis out of the realm of computability and into computational complexity. This thesis says that any function computable by a physical device can be computed by a Turing-equivalent computer with at most polynomial slowdown. This 'Extended Church-Turing thesis', as Shor (1988) calls it, might be

false. If human nervous systems are counterexamples to this thesis, then for sufficiently long instances of the Turing test, our best simulations will eventually crawl to a halt.

If these modified Church-Turing theses are true, and the contingencies don't prevent us from modelling brains and nervous systems and relevant environments, then it seems that the counterexample I have offered to French's claim has real world possibility. French's claim, when understood in its presented strong modal sense is clearly false. We can modify his claim to become merely one about this world, namely that no machine we can actually build will pass the Turing test by failing subcognitive questioning. To justify this will require more than listing off questions which appeal to embodied experience. It will require either offering reasons to think that one of the variants of the Church-Turing thesis is false for relevant reasons. Or, it will involve discussing relevant contingencies which prevent us from accomplishing an impractical, nomological possibility.

I do not have an answer here as to whether the physical or extended Church-Turing thesis is true. I have argued previously that there are not, as yet, any convincing refutations of the possibility of building a machine that outcomputes a standard Turing machine (citation suppressed). As for the extended Church-Turing thesis, there has been considerably less interest in the philosophy of computer science in investigating whether physical objects might display complexity properties that standard computers don't, prompting one researcher to comment that there is "...an overemphasis on the concept of computability"' (Volchan, 2002, 61).

The in-principle possibility of building a machine that passes the Turing test in the manner described would follow from the conjunction of the physical and extended Church-Turing theses. Notice, though, that the failure of either for biological systems, or other parts of the Turing test arrangement, could render the putative machine I have described impossible. First, notice that any finite function is Turing computable, and therefore physically computable. Second, notice that an emulation of any finitely specified performance can be achieved: load all responses into a finite amount of random access memory, and have a clock provide desired delays. A central characteristic of the Turing test is its unbounded nature. As many have noticed, no preset, finite bound on the length of the test can be specified before conducting the test; doing so renders the test trivial (an early expression of this can be found in Sampson (1973)).

The philosophical discussion of the Church-Turing thesis, and related forms, is vast. Historical questions, such as whether Alan Turing or Alonzo Church intended their theses to apply to physical objects, for example, is disputed (a useful summary can be found in Copeland (2002)). Descriptions of machines that compute functions no Turing machine can compute, the existence of which are consistent with known physics, have been offered (many of these are summarized and critically evaluated in Shagrir and Pitowsky (2003); Shagrir (2004)). There may be no answer to the truth or falsity of these theses forthcoming. However, I hope to have shown that no argument that ignores these theses can convince us of whether or not a machine can be built that passes the Turing test and that, according to both Turing and Descartes, deserves to be considered a thinking thing.

Notes

[1]Due to copyright restrictions governing the systems emulated, websites offering such services come and go quickly. For a selection of such sites, one can Google the terms 'web based emulator.'

[2]For example, consider work supporting the so-called 'enactive theory of perception' as presented, for example, in (Noë, 2004).

[3]This reply relies on my own views on answers to presently open questions in the philosophy of computer science, identified by Turner and Eden (2009).

[4]I have avoided in this paper discussion of computational functionalism: the theory that having a mind is nothing more than implementing a particular computer. If true, then philosophy of mind can be *identified* with the philosophy of computer science.

References

Copeland, B. J. (Fall 2002). The Church-Turing Thesis. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Cottingham, J. (1992). Cartesian dualism: theology, metaphysics, and science. In Cottingham, J., editor, *The Cambridge Companion to Descartes*. Cambridge University Press.

Cummins, R., Blackmon, J., Byrd, D., Poirier, P., Roth, M., and Schwarz, G. (2001). Systematicity and the cognition of structured domains. *Journal of Philosophy*, 98:167–185.

Deutsch, D. (1985). Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London A*, 400:97–117.

French, R. (1990). Subcognition and the limits of the Turing Test. *Mind*, 99(393):53–65.

French, R. M. (2000). Peeking behind the screen: the unsuspected power of the standard Turing Test. *J. Exp. Theor. Artif. Intell.*, 12(3):331–340.

Noë, A. (2004). *Action in Perception*. MIT Press.

Sampson, G. (1973). In defence of Turing. *Mind*, 82(328):592–594.

Shagrir, O. (2004). Super-tasks, accelerating turing machines and uncomputability. *Theoretical Computer Science*, (317).

Shagrir, O. and Pitowsky, I. (2003). Physical hypercomputation and the church-turing thesis. *Minds and Machines*, 13:87–101.

Shor, P. W. (1988). Quantum computing. *Documenta Mathematica - Extra Volume - Proceedings of the International Congress of Mathematicians*, I:467–486.

Turner, R. and Eden, A. (2009). The philosophy of computer science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2009 edition.

Volchan, S. B. (2002). What is a random sequence? *Mathematical Association of America Monthly*, 109:46–63.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Demo: Have a Chat with Sensitive Artificial Listeners

Marc Schröder, Sathish Pammi[1], Roddy Cowie, Gary McKeown[2], Hatice Gunes, Maja Pantic, Michel Valstar[3], Dirk Heylen, Mark ter Maat[4], Florian Eyben, Björn Schuller, Martin Wöllmer[5], Elisabetta Bevacqua, Catherine Pelachaud, Etienne de Sevin[6]

Abstract. Sensitive Artificial Listeners (SAL) are virtual dialogue partners based on audiovisual analysis and synthesis. Despite their very limited verbal understanding, they intend to engage the user in a conversation by paying attention to the user's emotions and non-verbal expressions. The SAL characters have their own emotionally defined personality, and attempt to drag the user towards their dominant emotion through a combination of verbal and non-verbal expression. The demonstrator shows the publicly available, fully autonomous SAL system.

1 INTRODUCTION

The SAL demo supports sustained emotionally-coloured machine-human interaction using non-verbal expression. The system aims to engage the user in a dialogue by paying attention to the user's non-verbal expression, and reacting accordingly. It focuses on the "soft skills" that humans naturally use to keep a conversation alive.

To simplify the challenge somewhat, the SAL system avoids task-oriented dialogue. Instead, it models the type of interaction found at parties: you listen to someone you want to chat with, and without really understanding much of what they are saying, you exhibit all the signs that are needed for them to continue talking to you. Similarly to the Rapport agent [2], SAL characters show non-verbal listener signals; in addition, they can also speak to engage the user in a simple dialogue. The approach has been test-run using Wizard of Oz setups at various stages of maturity [1]. This has allowed us to fine-tune the scripts used by the various characters, in order to react to the emotional state of the user in plausible ways despite the lack of verbal knowledge. The system is freely available from the project website `www.semaine-project.eu`.

2 THE DEMONSTRATION SETUP

In the SEMAINE demonstrator, one human user is sitting in front of a computer screen showing the face of an Embodied Conversational Agent (ECA). The user is wearing a headset for voice analysis and is recorded by a video camera for facial expression analysis. The ECA is speaking through loudspeakers, and is showing both verbal and non-verbal behaviour. A second screen shows a system monitor, displaying graphically the current information flow in the system.

[1] DFKI GmbH, Saarbrücken, Germany, email: firstname.lastname@dfki.de
[2] Queen's University Belfast, UK, email: (r.cowie | g.mckeown)@qub.ac.uk
[3] Imperial College London, UK, email: firstname.lastname@imperial.ac.uk
[4] Universiteit Twente, The Netherlands, email: (d.k.j.heylen |maatm)@ewi.utwente.nl
[5] Technische Universität München, Germany, email: lastname@tum.de
[6] CNRS, Telecom Paristech, France, email: firstname.lastname@telecom-paristech.fr

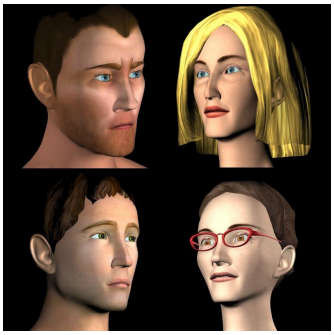

Figure 1. The four SAL characters represent the four quadrants of arousal-valence space: Spike is aggressive; Poppy is cheerful; Obadiah is gloomy; and Prudence is pragmatic.

The user can speak to one of the four SAL characters at a time (see Figure 1). Each character will try to sustain the conversation by being an active speaker and listener using multimodal verbal utterances and feedback signals. After a while, the user can switch to a different character. In the lab, sessions typically last for around 20 minutes; in the demo, much shorter sessions with changing users are anticipated.

Technically, the demonstrator system is a multimodal interactive system with components integrated across programming languages and operating systems by means of a standards-based framework for building emotion-oriented systems, the SEMAINE API [3]. Details on the technological setup and the individual processing components are described in a set of project deliverable reports available from the project website.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 211486 (SEMAINE).

REFERENCES

[1] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, 'The sensitive artificial listener: an induction technique for generating emotionally coloured conversation', in *LREC2008 Workshop on Corpora for Research on Emotion and Affect*, pp. 1–4, Marrakech, Morocco, (2008).
[2] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, 'Creating rapport with virtual agents', in *Intelligent Virtual Agents*, 125–138, (2007). doi:10.1007/978-3-540-74997-4_12.
[3] M. Schröder, 'The SEMAINE API: towards a standards-based framework for building emotion-oriented systems', *Advances in Human-Computer Interaction*, **2009**(319406), (2009). doi:10.1155/2009/319406.

Proceedings of the Towards a Comprehensive Intelligence Test (TCIT): Reconsidering the Turing Test for the 21st Century Symposium,
Aladdin Ayesh, Mark Bishop, Luciano Floridi, and Kevin Warwick (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

A proposed replacement for the Turing Test

Douglas A. Samuelson, InfoLogix, Inc., Annandale, VA USA

Abstract. As the Turing Test of intelligence relies on deception, we seek a more general test of intelligence that includes the Turing Test as a special case. We posit that essential characteristics include meta-learning (learning about learning), meta-generation (generation of rule generators), meta-assessment (assessment of assessment methods), and awareness of other beings' emotional states and reactions as well as their logical processes. Accordingly, in this brief position paper, we offer a new proposed test: An intelligent entity can learn how its actions increase or decrease other sentient beings' satisfaction or dissatisfaction with their current state and circumstances, and can then manipulate that information to its advantage

1 INTRODUCTION

The Turing Test of machine intelligence is brilliantly simple. When proposed, it represented a major improvement over more complicated theories that required complex philosophical assumptions. It is both logically solid and intuitively appealing: if the machine can somehow learn what abilities to display and not to display to convince a human experimenter that it strongly resembles a human, then it is, indeed, an entity most people would consider intelligent.

There are two problems with this definition. First, it is based on deception. This makes some of us uncomfortable, as we would like to consider as well a machine that can please us honestly. This is a special case of the second problem: the Turing definition is perhaps sufficient but clearly not necessary. Without making the assessment hopelessly complicated, we would like to broaden the test to include other ways to demonstrate intelligence.

2 PROBLEM STATEMENT

We seek a test for machine intelligence that is both practical and exhaustive: it should be both necessary and sufficient as a criterion for machine intelligence. The Turing Test fails the "necessary" criterion, as we can conceive of clearly intelligent non-human entities that would not pass the Turing Test.

In this author's view, the new test should:
-- incorporate learning, generation of rules, generation and assessment of generators;
-- incorporate awareness of others' emotions, where present, not just logic;
-- not be species-specific;
-- take into account that an intelligent entity may not be biological or electromechanical: organizations count too (for example, Hofstadter's "Aunt Sarah," the ant colony from *Goedel, Escher, Bach*);
-- avoid the "Searle's Chinese Room" paradox in which entities can appear to behave intelligently by following a script but lack other essential features, such as ability to adapt;
-- not require frequent human intervention and assistance to achieve and maintain high performance;
-- imply the Turing Test as a special case; and
-- remain both practical to apply and theoretically sound.

3 DISCUSSION

Recent research indicates that learning and adaptation are essential elements of our definition. Moreover, learning about learning, assessing means of assessment, and adapting ways of adapting – in short, recursive layered behaviors – now appear to be critical components of intelligence.

One counter-argument to the Chinese Room paradox is that it depends on an artificial constraint: it relies on the assumption that the experimenter not only cannot distinguish between the translations produced, but also cannot ask the translators to do so. In real life, asking experts to critique each other's expertise is common practice, and often most informative. Similarly, in the Turing Test situation, the experimenter could ask each entity to comment about whether the other seems human. The entities would be challenged not only to learn how to present their abilities to a human judge, but also to learn why they should try. That is, their rewards for trying to pass the test would not only influence their behavior in the test but also teach them that the experimenter wants to be deceived; we might then wish to teach them when deception is not desired. Thus we would place greater value on the machine that can not only learn to deceive us but also learn when not to do so.

This implies, in turn, an ability to learn not just how to solve the problem presented, nor even just how *not* to solve sub-problems in a way that gives away information better concealed, but also to learn about the experimenter's likes and dislikes. Just as with other people, we are likely to regard most highly not the fastest logical processor, but the entity that can most quickly learn to assess other sentient beings' emotional reactions to its actions and adjust its behavior accordingly, to its advantage.

We therefore come to the following proposed test.

4 PROPOSAL

To be considered intelligent, the entity demonstrates that it can learn how its actions increase or decrease other sentient beings' satisfaction or dissatisfaction with their current state and circumstances, and can then manipulate that information to its advantage.