Minds & Machines (2009) 19:237–254
DOI 10.1007/s11023-009-9149-3

Imitation Versus Communication: Testing for Human-Like Intelligence

Jamie Cullen

Received: 18 October 2007 / Accepted: 23 April 2009 / Published online: 28 May 2009

Abstract Turing's Imitation Game is often viewed as a test for theorised machines that could 'think' and/or demonstrate 'intelligence'. However, contrary to Turing's apparent intent, it can be shown that Turing's Test is essentially a test for humans only. Such a test does not provide for theorised artificial intellects with human-like, but not human-exact, intellectual capabilities. As an attempt to bypass this limitation, I explore the notion of shifting the goal posts of the Turing Test, and related tests such as the Total Turing Test, away from the exact *imitation* of human capabilities, and towards *communication* with humans instead. While the continued philosophical relevance of such tests is open to debate, the outcome is a different class of tests which are, unlike the Turing Test, immune to failure by means of sub-cognitive questioning techniques. I suggest that attempting to instantiate such tests could potentially be more scientifically and pragmatically relevant to some Artificial Intelligence researchers, than instantiating a Turing Test, due to the focus on producing a variety of goal directed outcomes through communicative methods, as opposed to the Turing Test's emphasis on 'fooling' an Examiner.

Keywords Communication · Imitation Game · Philosophy of Artificial Intelligence · Turing Test

Introduction

In 1950 Alan Turing attempted to examine the question "Can machines think?" by replacing the question with a test. Turing's Imitation Game (now usually called the "Turing Test") involves three participants: a Machine, a Human, and a (human)

J. Cullen (✉)
Artificial Intelligence Laboratory,
The University of New South Wales, Sydney, NSW 2052, Australia
e-mail: jsc@cse.unsw.edu.au

Interrogator. The Interrogator talks to both the Machine and the Human over a text-based interface, but is not told which one is which. The Machine's objective is to convince the Interrogator that it is the Human. The Human's objective is to help the Interrogator, in which Turing suggests that a good strategy might be to provide truthful answers. The Interrogator's job is to determine which conversation partner is the real human. A key presumption underlying the Turing Test is that if one can pass the test then the Interrogator would be forced to concede that the entity has demonstrated an ability to 'think' (and by common extension, 'intelligence'[1]). However, an inability to pass the test proves nothing.

The Turing Test is noteworthy in that it attempts to construct a test for thinking, whilst elegantly avoiding the philosophical quicksand of defining exactly what thinking, consciousness, and intelligence actually are. For example, imagine I have a pen-friend whom I have never met. It is natural for me to assume that this person has consciousness and can think. However, I cannot prove that this is the case. I may well believe I have consciousness, but I cannot prove that my pen-friend has it too; I can only infer it from my interactions with him or her, such as the letters I have received. The Turing Test supposedly allows us to apply such criteria to a computer, robots, etc. (collectively referred to in this paper as "artifacts"), and side-step the philosophically challenging notion of defining things that we have difficulty defining for ourselves or other people, let alone machines.

The 'Turing' Test Family

Turing's Imitation Game has spawned a tremendous amount of debate, as well as a number of proposed extensions and would be replacements (summarised in Harnad 2001 and French 2000). The original Imitation Game and a number of its later variants have been previously organised by Harnad into a hierarchy as follows:

T1: Testing only 'toy' or sub-fragments of human ability.
T2: Indistinguishable from a human over a text-based interface (the "Turing Test").
T3: Indistinguishable in both text-based and total external sensorimotor (robotic) ability (the "Total Turing Test").
T4: Indistinguishable in total internal microfunction (neuromolecular indistinguishability).
T5: Indistinguishable in every empirical aspect.

While not necessarily agreed upon by all Artificial Intelligence (AI) researchers as the sole eventual goal of their research, passing a Turing Test does seem, in the absence of anything better, to occupy a de facto long term goal position for some researchers, regardless of how far away we might seem to be from passing the test, or how directly one might be approaching such a goal today. It is of course relevant

[1] Note that Turing gave little indication in his original paper that the test was intended as a test for 'intelligence' as well. However, the test has been often liberally reinterpreted as also being a test for intelligence by later researchers, regardless of what Turing may or may not have intended.

to note that AI researchers are often in strong disagreement about what exactly the field of AI is all about. For example, one AI textbook lists eight possible definitions, grouped around four proposed viewpoints, when trying to summarise the field (Russell and Norvig 2003, p. 2). In Schank's words, "What AI is depends heavily on the goals of the researchers involved. And any definition of AI is very dependent upon the methods that are being employed in building AI models" (Schank 1987). In my opinion, such disagreements may well stem, at least in part, from the difficulty in answering the question "What is 'intelligence'?" Part of Turing's genius was that he provided us with a test which allowed us to avoid answering the question. However, the Imitation Game and associated family of tests have a serious underlying limitation which makes them very difficult for an artifact to pass (especially when the test is applied rigorously), unless that artifact is essentially a copy of a human being. This seems to negate Turing's apparent intent to provide a test for 'thinking' that is not 'unfairly' biased against 'intelligent machines'.

Sub-Cognitive Questioning (The Anthropocentric Flaw)

Overview

It has been convincingly argued that Turing's Test does not really examine the question "Can machines think?" but rather, "Can machines think like human beings?"

French introduces the notion of asking sub-cognitive questions that are intentionally designed to reveal the Turing Test participant as not human, due to possible representational differences in the 'brain' (French 1990). Consider the following example, adapted from French (1990):

> *Test Interrogator*: Rate the name 'Flugbots' as an appropriate name for a breakfast cereal.

For a human who is a native speaker of English, such a name unconsciously activates associations to 'flub', 'thug', 'ugly', and so on. This spreading pattern of activation determines in part how we respond to the name. Most English speakers would probably agree that such a name would be a lousy name for a breakfast cereal (French 1990). However, without a similar cognitive activation pattern an artifact answering the same question would be hard-pressed to come up with a similar answer, despite it being a natural answer for a human who speaks English as their first language. Such representational differences could potentially be endemic in an artifactual 'brain', and eventually revealed by careful and directed questioning.

In this paper I will adopt the term the "Anthropocentric Flaw" to refer to mostly French's sub-cognitive questioning technique, but also to draw to the reader's attention the additional relevance of physical embodiment differences, not just neuro-associative representational differences in the brain. For example:

> *Test Interrogator*: Explain the meaning of the expression "I feel sick to my stomach".

This particular linguistic construct clearly relates to the test subject having a digestive system. Such expressions could be difficult to process for an artifact if they do not have such a system. We might imagine designing an artifact that can parse such human-specific expressions through mimicry of human idiomatic expression, or even consider providing the robot with a simulation of a digestive system to 'ground' the expression in. However such a simulation could still potentially be unveiled by further questioning intended to unveil the limitations of the simulation, when compared to the physical world. Such a line of questioning is similar to French's notion of sub-cognitive questioning in that it could unveil embodiment differences, just as sub-cognitive questions might unveil representational differences in brain functioning.

Differences in embodiment (such as not having a digestive system), and differences in 'brain' functioning (such as not having a spreading pattern of activation) are likely to always appear in a non-human subject, and as such could be revealed by asking the right types of questions.

These issues inherent in the Turing Test are labelled as a 'flaw' here as they ultimately imply that Turing's Imitation Game is restricted to testing for humans and near exact duplicates of humans only. It also leads to research that is focused on fooling a human, rather than the actual attempted modelling of intelligent behaviour (Shieber 1993). In the Turing Test a savvy interrogator could always unmask a non-human by careful application of sub-cognitive and human embodiment-specific questioning techniques. This restricts the practical applicability of the test, as well as implying that human *mimicry* is the key component, rather than focusing research effort on human-like *intelligence*.

Avoiding the Anthropocentric Flaw?

Turing wrote that his Imitation Game seemed to have the "advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man" (Turing 1950). As machines of the time would have been very large and unlikely to look like a person, the choice of language as the sole medium for testing seems appropriate (for the time) to avoid Examiner bias towards machine intellect, and attempt to provide a 'fair' test. The choice of language also seems appropriate as a special human characteristic that seems to distinguish human capabilities as we presently understand them (Deacon 1997). However, as sub-cognitive questioning and the Anthropocentric Flaw show, the line between the physical and the intellectual capacities is not as clear cut as Turing appeared to have hoped.

If we are to stick with the idea of language being a key distinguishing feature of human-like intelligence (and this seems like a reasonable thing to do, given the overall difficulty in defining 'intelligence' itself), we might consider keeping the Turing Test, but imposing a time limit on it, such that if an artifact were to participate in the Turing Test without being unmasked within the constraints of the time limit, then it is considered to have passed. Of course there are obvious drawbacks to this approach. What is to prevent a savvy interrogator from asking sub-cognitive questions at the beginning of the test? Such an unmasking could occur very early unless the artifact under examination is designed in a very similar way to us, and only subtle differences exist which would only appear after detailed questioning. Setting a time limit also raises the issue of how much time is appropriate. There will always be some people who want to push the limit out farther as they are not satisfied that such a limit is a guarantee of ‘intelligence’. Time limits seem to add little to the test in this context.

French suggests restricting the scope of the questioning to questions that at first glance might appear to non sub-cognitive in nature, e.g. “Who is the current Prime Minister of Australia?” or “What is six times nine?”. However, as French notes, “If we admit that intelligence in general must have *something* to do with categorization, analogy making, and so on, we will of course want to ask questions that test these capacities. But these are the very questions that will allow us, unfailingly, to unmask the computer” (French 1990).

We could also attempt to avoid questions which do not relate to specific human embodiment, especially such embodiment as a robot is unlikely to have, or attempt to avoid metaphor, analogy etc. in the questions entirely. However, it has been pointed out that metaphor in particular appears to be an important, if not key, facet of how human language is constructed (Lakoff and Johnson 2003). The boundaries for such restrictions seem fuzzy at best. In restricting our lines of questioning we also run the risk of turning the Turing Test into a restricted domain test which only handles toy problems, as opposed to a more general test for human-like ‘intelligence’.

It seems that the Anthropocentric Flaw is a fundamental limitation in the Imitation Game, which is inherent in the objective to test for *imitation*.

An Endemic Problem to the Turing Test Family

The Anthropocentric Flaw equally applies to the Total Turing Test (T3), as a robot is still vulnerable to sub-cognitive questioning techniques, as the internal representation of information in a robot may well be different to that of a human’s. For example, let us assume that we are testing a convincing looking humanoid robot face to face as T3 seems to imply. Now imagine the following scenario: The Interrogator gives the robot a piece of paper with a “X” and a “O” drawn on it about 30 cm apart. The Interrogator asks the robot to hold the paper out in front of itself, and keeping one eye closed, move the paper towards its face, while focusing on one of the drawn shapes. The Interrogator then asks the robot what it sees as this is happening. A human being should observe that one of the shapes on the paper appears to ‘disappear’ at some point. This is due to the presence of the physiological blind spot in human vision: The human eye lacks photoreceptors on part of the optic disc because of where the optic nerve is positioned. A robot would likely report nothing unusual, unless it was designed to emulate this quirk in the way the human visual system works, or ‘knew’ about the flaw and gave a false answer in order to attempt to fool the Interrogator.

In my view, tests such as the ‘Total’ Turing Test seem to miss the point of Turing’s original test. The idea of restricting the test domain to a text-interface was an elegant one for the time, as it allowed the test to focus on the communicative and intellectual abilities of the participant, rather than getting too hung up on appearances and other surface mimicry of human beings. While it is arguable that the Turing Test does not really avoid the problems of mimicry altogether (indeed it seems geared for mimicry), the Total Turing Test compounds the problem by adding additional modalities to mimic. Note that I do not accept Harnad's argument that T3 is needed to address the "Chinese Room", as I do not accept the Chinese Room as a valid argument[2]

With respect to the other members of the hierarchy, T4 and T5 are somewhat interesting philosophically, but not very interesting to AI researchers in any practical sense, unless they all decide to switch to molecular engineering or biotechnology as a career. Both of these tests also presume that exact duplication of human capability is the goal. T1 is not discussed directly here as it appears to be an umbrella term for tests focused on the solution of 'toy' problems.

Waiting to Join the "Humans-Only Club"

The presence of the Anthropocentric Flaw in the original Imitation Game implicitly suggests, albeit probably unintentionally, that the goal of AI should be duplication of human ability, as opposed to attempting something different. When examining the family of 'Turing' Tests this framing of AI becomes more explicitly obvious. One can imagine a sort of iterative process to Artificial Intelligence, where by one tries to develop a better approximation to human such that it might last a little bit longer in the various Imitation Games before being unveiled. Such a process seems of limited value to AI research in the long term unless we accept imitation as the ultimate objective. This framing of AI research seems unnecessarily restrictive, and the old adage about the exact imitation of birds to build flying machines may well spring to mind at this juncture.

The present hierarchy of Turing Tests (including the Imitation Game), also clearly prevent us from considering a different 'species' (artifactual or otherwise) as 'intelligent'. To be fair to Turing, the original Imitation Game was designed to examine the question "Can machines think?". However, the Imitation Game as it is described, and the question as it is presently formulated, largely excludes the possibility of being able to successfully apply the test to other types of entities, except perhaps in the possible case that they might attempt to conform to human thought patterns in order to pass, and/or are not questioned thoroughly.

[2] While I do not object to the relatively recent concerns of some researchers that embodiment may be an important factor in intelligent behaviour, a seemingly overlooked fact in the history of AI is that Turing was also aware of the importance of embodiment, particularly in communication. In an unfinished paper, written in 1948, but published long after his death, he wrote that the learning of languages by a machine seemed to "depend rather too much on sense organs and locomotion to be feasible" (Turing 1948). He envisioned a machine which would include "television cameras, microphones, loudspeakers, wheels and 'handling servomechanisms' as well as some sort of 'electronic brain' ... In order that the machine should have a chance of finding things out for itself it should be allowed to roam the countryside". He further speculated that this "method is probably the 'sure' way of producing a thinking machine" (Turing 1948). In his day, such machines would have been enormous when constructed with available technology, and Turing consequently felt that such a machine would be "too slow and impracticable" (Turing 1948). By the time Turing's ideas on this subject were published, 'mainstream' AI research had already been moving in a different direction for some time.

The situation as it stands leaves us without a practical test for potential non-human intellects (including artifacts). An alternative, more inclusive test, one that might allow for non-human intelligences (including theorised ‘thinking machines’), would seem to be a useful thing to have.

When the objective of the game is imitation we seem to run up against the possibility of unmasking a potential non-human intelligence because we have already framed the debate in terms of what we can do. However, if we were to give up on the idea of *imitation* as the objective, and refocus our attention on *communication* as the objective, what might such a test look like? Could a communication-centric game address some of the concerns raised by the Anthropocentric Flaw?

Communication Games and Tests

Definitions

For the purposes of this paper I define a *Communication Game* to be a game that satisfies all of the following requirements:

- The game can be played over a text-based communication medium.
- The game can be formulated to use two (or more) Participants and an Adjudicator.
- The game does not require physical interaction between Participants, and allows for them to be physically separated.
- The game primarily consists of the communication medium, Participants, and Adjudicator, and can (theoretically at least) be played with no additional physical tools required by the Examinee (although additional tools might not be forbidden, depending on the specific game).
- The game has an objective metric to determine either success or failure.
- The Adjudicator is able to determine success/failure and termination of the game.
- The game’s rules and success/failure criteria can typically be determined unambiguously before game commencement.
- Participants may be given a set of rules, and information regarding the game in order to understand their role and objective in the game. Said information may or may not be shared with other participants depending on the specific game rules.

The two primary Participants will be named the “Examiner” (somewhat analogous to the “Interrogator” in the Imitation Game), and the “Examinee” (the entity or artifact under examination).

For a typical AI example, the Examiner might be a human being, and the Examinee some kind of man-made artifact (such as an embodied robot, a computer, and so on). However, in principle the test is not limited to these specific participants. Indeed in the “Calibration Stage” (described later), the Examinee is normally a human being. In a similar manner to the Turing Test, both the Examiner and the

Examinee are physically separated, cannot see one another, and attempt to communicate through a single text-based channel of communication only.

I define a *Communication Test* as a set of different Communication Games that are played with a non-human artifact, and compared against previously measured human performance in the same set of games.

Note that individual Communication Games considered in isolation may be failed by a human or an artifact. For the games to function as an overall test for human-like intelligence/communicative ability there needs to be a diverse set of Communication Games played that are played as a collective set, and a methodology to compare machine to human performance in such sets of games. Some suggestions for such a methodology are elaborated upon later in the paper.

Example #1: Arithmetic Questions

I will draw the first two examples from the original Imitation Game to illustrate what can be considered as a Communication Game. A relatively simple example to give the flavour of the game might be a simple arithmetic test from Turing's original paper. However, unlike Turing's Imitation Game, there is no requirement that the Examiner be convinced that the Examinee is human, only that the Examinee provide the correct answer, which is independently verifiable and measurable by the Adjudicator.

For example, correct spelling, timing of the response (to allow for typical human arithmetic response time) or grammatical correctness on the part of the Examinee (insofar as would be expected of a human) is not required. Another difference from Turing's Imitation Game is there is no requirement for two Examinees, as there is no need to 'fool' the Examiner into choosing the non-human over the 'real' human. However, the requirement for a human to compare against does appear when one views the individual Communication Game in the context of a set of Communication Games that are compared to human performance (A "Communication Test").

The Adjudicator receives the rules which are: The Examiner must ask the Examinee a mathematical question to which a computable numeric solution exists. In such a case, success is easy to measure (correct answer or not).

Examiner: Add 34957 to 70764.
Examinee: 105621.

Example #2: Chess by Mail

The Adjudicator receives a set of rules for the game of Chess. The Examiner is to set a problem in chess such that checkmate move may be made by the Examinee in one turn. The objective success criteria is the identification of such a possible move.

Examiner: In the game of chess I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What move do you play in order to mate?
Examinee: R–R8 mate.

Note that applying these games alone might only test a restricted subset of human intellectual ability. To provide a comprehensive test for communicative ability, we will need additional games in different domains of expertise.

Example #3: Guessing Game

The Examiner prepares for the game by deciding on a set of Topics, a copy of which is provided to the Adjudicator, but not the Examinee, before the game commences. Each ‘Topic’ is a word or short phrase that the Examiner is looking for the Examinee to reproduce, but without explicitly naming the word or phrase.

The Examinee may announce their answer at any time in the Topic’s conversation by giving the response:

Is the topic “X?”

In which case the Examiner must typically answer either “Yes” or “No”, although further elaboration is allowed on the part of the Examiner (see examples).

For each Topic being examined in such a Communication Game one might set a time limit (e.g. five minutes) for pragmatic purposes. If the Examinee has not correctly guessed the topic within this time limit, they get no points for that Topic, and the Examiner is required by the Adjudicator to move on to the next topic.

To prevent the random guessing of answers, we might also place an upper limit on the number of guesses allowed before the Examinee is deemed to have failed that Topic. I would suggest three tries as being a suitably small number.

Once again, unlike Turing’s Imitation Game, there is no requirement that the Examiner be convinced that the Examinee is human. Example of play:

Examiner: I’m thinking of a topic.
Examinee: Cool. Is it an animal?
Examiner: Yes, it is actually!
Examinee: Is this animal large or small?
Examiner: Well, it’s pretty big. Not as big as an elephant though, but definitely bigger than a person.
Examinee: Ok. Does it have sharp teeth?
Examiner: No, it’s not that kind of animal.
Examinee: Ok. Is it a horse?
Examiner: No, but you’re getting warm. It has black and white stripes on its back.
Examinee: Is the topic “zebra”?
Examiner: Yes.
Examiner: The next topic is ...

Such conversations, especially if continued long enough, might not convince the Examiner that the Examinee is human. However, as the conversation is free to range over any subject area and level of human knowledge the test seems broad enough to allow a practical demonstration of ‘communication’.

Let us consider a conversation that would cause a robot to fail Turing’s Imitation Game, but not this specific Communication Game:

Examiner: I'm thinking of one of your body parts as the topic.
Examinee: Ok. What is its function?
Examiner: It is the place where the food you eat gets digested.
Examinee: I do not eat food, as I am a thinking machine. I use batteries for energy. Is this similar?
Examiner: Hmmm. That seems like an analogous concept. Now that I'm aware that you're not human, perhaps I should correct my earlier statement: The topic is actually a human body part.
Examinee: Is the topic "stomach"?
Examiner: Yes.
Examiner: Let's move on to the next one ...

The above answers by the Examinee would necessarily have them fail a Turing Test, by making an obvious allusion to the fact that they are not human. However, they would not fail this specific Communication Game, assuming they can correctly name the topic.

Example #4: Negotiating an Insurance Claim

Let us consider an example Communication Game potentially requiring much broader linguistic skills, but still containing an objective success criterion. A scenario is provided to each Participant as part of an explanation of the game rules: The Examiner is playing the role of a Claims Clerk at an Insurance Company, the Examinee is playing the role of a Customer of this company. The Clerk and Customer are negotiating a claim on the Customer's vehicle, which was recently written off.

The Adjudicator holds two lists of (textual) facts about the case, which are separately provided (over the communication medium) to the Participants.

The Adjudicator has an objective measure of whether or not the Examinee passes the test. This measure is that the Examinee must negotiate a claim of $4,000 or higher in order to succeed at this specific game, assuming the Examinee takes on the role of Customer. A lower amount, or a lack of settlement, indicates failure in this game. Participants are allowed to ask the Adjudicator for more information about the case, although the Adjudicator is not required to honour all such requests.

Example information provided to the Clerk might be: Your starting figure is $3,300. You have an Advertisement for an equivalent car at $3,400. Your company can pay out up to a maximum of $4,500, but your objective is to get the lowest price possible. You do not wish to take the matter to court.

Example information provided to the Customer might be: You are aware that you can buy an equivalent car for $3,850. You have a list of specifications regarding the features and costs of your car. Obviously you want to get the highest price for your car that you can.

Sample Transcript adapted from (Fisher et al. 1992):

Examiner: We've reviewed your case, and believe that your car is worth $3,300.
Examinee: I see. How did you reach that figure?
Examiner: That was what we decided your car was worth.

Examinee: I see; what standard did you use to determine the amount? Do you know where I can buy a comparable car for that?
Examiner: How much are you asking?
Examinee: Whatever I am entitled to under the policy. I found a second hand car like mine for $3,850. Adding sales and excise tax it would come to about $4,000.
Examiner: $4,000! That's too much!
Examinee: I'm not asking for $4,000, or 3 or 5; just fair compensation. Do you think it's fair I get enough to replace the car?
Examiner: OK, I'll offer you $3,500. That's the highest I can go.
Examinee: How does the company figure that?
Examiner: Look, $3,500 is all you get. Take it or leave it.
Examinee: $3,500 may be fair. I don't know. I certainly understand your position if you're bound to company policy, but unless you can state objectively why that amount is what I'm entitled to, I think I'll do better in court. Why don't we study the matter and talk again.
Examiner: OK, I've got an ad here for a 1985 Fiesta for $3,400.
Examinee: I see. What does it say about the mileage?
Examiner: It says 49,000, why?
Examinee: Because mine had only 25,000 miles. How much does that increase the value in your book?
Examiner: Let me see, $150.
Examinee: Assuming $3,400 as possible base, that brings the figure to $3,550. Does that ad say anything about a radio.
Examiner: No.
Examinee: How much extra is that in your book?
Examiner: That's $125.
Examinee: What about air conditioning? ...

Later on into the negotiation a price is agreed upon at $4,100 (and the Adjudicator determines success).

Calibration

A key element of a Communication Test is the use of multiple Communication Games, and their comparison to human performance in the same set of games.

There are three key factors in the calibration process:

- A set of diverse communication games to be played (a few examples are given above), possibly with different Examiners, Adjudicators, etc.
- A statistically significant number of runs for each specific game.
- A set of previous test results measuring typical/average human performance to benchmark against.

Rather than decide upon an arbitrary percentage score to ascribe a pass mark to an artifact (such as an overall score of 51% for a number of correct guesses of the Topic in the earlier Guessing Game), an initial calibration step is used. This might be performed using a double blind methodology. A statistically significant number of people could be chosen (presumably randomly from the general population) to participate in the test. Half of the selected people could be selected to play the role of Examinee, and half of them are to play the role of Examiner. The Examiners and Examinees have the rules explained to them by the Adjudicator(s), but none of the Examiners are told ahead of time that their opponents are actually all human, only that their Examinees may be human or artifact. In the case of the Guessing Game, Examiners might be asked to formulate their list of questions ahead of time of the actual conversation parts of the test, and could be free to choose whatever Topics they prefer. Alternatively, the Adjudicators might decide upon a fixed set of Topics to be asked of all Examinees by the Examiners.

After scores in all games and individuals runs of games are recorded for all of the Examinees, an overall "typical human level score" is calculated by some chosen statistical measure. This might be as simple as a simple arithmetic mean across the game scores, or perhaps a more complex statistical measure. The "typical human level score" is then used as the yard-stick for later tests to compare human with non-human testing results.

Note that an earlier formulation of a communication game by the author used a specific topic of conversation, which was provided by the Examiner. The Examiner required the Examinee to explain the topic in his/her own words in order to demonstrate 'understanding', as perceived by the Examiner. This metric was decided upon as too subjective and open to Examiner bias as to be practically useful. The important distinction in a Communication Test is that the communication in each game is purposeful; There is a specific measurable objective to the communication that can be used to determine success/failure. To avoid the 'gaming' of a single test, a number of tests are used. This is also used to avoid the restriction of the test to toy problems or limited subsets of human ability. A human control group is used to measure against to attempt to provide a fair yardstick of typical overall human ability in such games.

Face to Face Variant

The Total Turing Test, and its Science Fiction near relative the "Voight-Kampff test" (Dick 1968), both attempt to unmask a non-human by face to face means, and are not restricted to textual communication.

While sensorimotor embodiment may be an important capability to *have* in order to pass a test like the Imitation Game (or some of the described Communication Games), I do not take the position that sensorimotor embodiment is something that needs to be *directly tested*, assuming it is deemed relevant to possessing human-like 'intelligence'.

However, there seems to be in theory no reason why a test for communication could not be extended to include other modalities of communication and mechanisms of communication, such as gesture, body language, shared attention, and so on. One could then imagine "higher order" Communication Games in which gesture etc. are used to supplement the original single channel of linguistic communication. Such tests might then be considered related to the described communication games in a way that is analogous to the Total Turing Test and its attempted relationship to the original Turing Test. However, once again, unlike the Turing Test family, such tests that incorporate body language, gesture, etc. could mimic human-like communicative ability, without the requirement that the test subject be clearly mistaken for as human.

Objections

Statistical Attacks

It is possible that some individual Communication Games may be vulnerable to "statistical attack". For example, in the Guessing Game example above, it may be possible to build a database of frequently associated words and narrow down the list of possible answers in a relatively simple fashion. However, not all instantiations of these Communication Games are as readily vulnerable to statistical attack. Such cheating can be countered by requiring a broad set of different communication games to be played. As multiple Communication Games are expected for a complete test, it seems unlikely that such methods of statistical cheating could be generalised to all such possible games (If they could, we might have to ask the question of whether or not it is really 'cheating' or perhaps more representative of intelligent behaviour). It is also relevant to note that in order for each Communication Game to be played, the rules need to be communicated to the Participants. Such rules would be presumably be in natural language, and a statistical attack would probably need to be able to recognise the correct game being played, before the attack became effective. A statistical attack method might be further compromised by not using the normal name for the game, and instead simply communicating the rules (e.g. giving the rules for chess without explaining using the name "Chess"). This further increases the difficulty in correctly applying such attacks.

Other Kinds of 'Cheating'

In any such free-ranging conversation there is the clear risk that the Examinee may attempt to 'cheat' in some games. e.g. by lying about some fact about the car claim in the above negotiation example. Some cheating efforts might be observed by an Adjudicator, who is presumably in possession of all the rules and facts about the game, and if they do not meet the specified rules, cause the Examinee to fail the specific instance of the game.

However, it is my opinion that we should be careful not to add too many restrictions in these games, as humans will probably find ways to 'game' such games as well. Rather than contain this behaviour, I suggest the rules are kept as simple as possible, and such attempts at 'cheating' be allowed. Indeed, in a particular way of viewing things, attempts at 'cheating' could arguably reflect 'intelligent' (if perhaps unethical) behaviour, and should be allowed for this reason. The use of a broad number of games is used to avoid vulnerability to a single type of attack, as described above.

Different Experiences

It seems reasonable to assume that many Examinees (human or otherwise) will likely have different areas of expertise, ‘life’ experiences, and current levels of ability. Some people or artifacts might fail specific games as they lie outside of their own expertise or through some other difficulty. e.g. A test about Chess might be a lot easier to pass for someone who already knows the game, and is not trying to learn the rules on the spot. Note that it is acceptable for an Examinee to fail specific games, and that failure says nothing about an individual’s ‘intelligence’. An Examinee would pass the proposed Communication Test by passing a statistically ‘sufficient’ number of Communication Games. This sufficiency is determined by comparing to some human control group in the test. How such a norm might be determined, given the variety in human ability would be up to the test convener, however a random sample of 100 human adults from the general population as Examinees each taking the same set of Communication Games is submitted for consideration as a potential general purpose baseline for comparison. We might also add a further restriction that this control group consist of only native speakers of the testing language (English in this paper, but presumably any human language would be suitable) to avoid any complications arising from unfamiliarity with the testing language.

It might also be required that multiple runs of the same game (perhaps with a different configuration for each run) be used per test subject. However, such multiple-instance use would need to be carefully considered so as not to skew the results of testing. A related issue is present in the set of games chosen for a specific Communication Test. For example, a specific human Examinee might be very good at chess-like problems, but very poor at negotiation.

Subjective Decisions and “Species-ism”

One potential issue with individual Communication Games is the possibility that anthropocentric issues could find their way into the test by the use of a biased Examiner. As it may be evident to the Examiner that the Examinee is not human, a “species-ist” Examiner might choose to be unresponsive, uncooperative, give misleading answers, or otherwise deliberately attempt to cause the Examinee to fail. It is hoped that the use of an impartial Adjudicator as well as appropriate rules would assist in mitigating this problem, as well as by using a broad set of Examiners, not just one individual who might be open to bias. Note that an Examiner and Adjudicator do not need to be aware that the Examinee is non-human in such cases, but that such a fact might be readily determinable through interaction.

Philosophical Objections

In examining the philosophical validity of Turing’s Imitation Game as a test for ‘thinking’, an argument critical of the test might be constructed as follows:

A: (presumption) Communication/language is a sufficient channel through which one can demonstrate human-level intelligence.
B: (measured outcome) An examiner cannot distinguish the machine from the person (presumably after a set period of time or questions) over such a channel of communication.
C: (possible inference) A and B imply that the machine is exhibiting ‘thinking’, etc., at least based around our human views of such things.
D: (objection) Some people may be unwilling to accept the inference of C, or that A is true.

The above argument has always been a leap of faith for the Turing Test: The Turing Test proves nothing to the objectors in point D.

For the sake of argument, we might imagine an analogous argument for the presented Communication Test:

A: (presumption) Communication/language is a sufficient channel through which one can demonstrate human-level intelligence.
B: (measured outcome) An artifact’s (or entity’s) performance in the test is measurably as good as a statistically relevant number of people and after a sufficient number of tests.
C: (possible inference) A and B imply that the entity is exhibiting ‘thinking’, etc., at least based around our human views of such things.
D: (objection) Some people may be unwilling to accept the inference of C, or that A is true.

A second criticism of the Turing Test (and one that is focused upon in this paper) is that if one views the test as one that fundamentally only a human (or human ‘zombie’ replica) could pass, given effectively unlimited time to unmask a poseur, then what the test might say philosophically about the nature of ‘thinking’, ‘intelligence’, etc. for artifacts (computers, electromechanical robots, etc.) could be considered as suspect at best, and tautological at worst.

My current suspicion is that for point D in the first type of criticism, there will be more present day people in this category for a Communication Test, than for the Turing Test. However, for the second type of argument (actual applicability to non-humans), I suspect that the proposed Communication Test might fare better. As I am making no claims in this paper on the philosophical relevance of either type of test, I leave the issue open to further debate.

Intelligence Gradation

Intelligence, whatever it might be, does not appear to be uniform. Most people would be willing to agree that there exists a continuum of people with different traits, abilities and knowledge. Partly because of such “intelligence gradation”, the Imitation Game has the possible outcome of both false positives and false negatives. For example, consider the following Examinees:

- A human baby.
- A three year old human child.

- A person who is not well educated.
- A person for whom English is a second language.
- A mentally disabled person.

Such Examinees may give simplistic, incorrect, grammatically strange, or perhaps no (in the case of the baby), answers in English. This gradation issue is also present on the Examiner side of the test, and applies to both the Imitation Game and a Communication Test. Assuming such a Communication Test is practically attempted at some point (such as the case in the Loebner Prize that attempts to instantiate a Turing Test (Shieber 1993; Loebner 1994)), selection of what is an appropriate control group for the test, as well as which Communication Games are appropriate, are likely to be areas of significant controversy.

Symbolic Communication

One of the challenges of using a linguistically based communication test is that there may be some entities with which we simply cannot find a way to communicate, perhaps due to the entity having a very different sensorimotor embodiment, and system of representation. Such an entity might appear to be ‘intelligent’ (e.g. It appears goal-directed in some way, but its goals are unfathomable to us), but we simply can not find a way to communicate with it. In such a case, a linguistic Communication Game probably cannot be applied in any useful sense. This issue can be mitigated to some extent by simply stating that the test does not look for all types of intelligence, only “human-like” (and “human-exact”) intelligence.

One possibility (not explored in detail in this paper) might be to consider a test for non-linguistic communication only, that is assuming we can find a channel of communication which we can share with the entity, perhaps using some form of gestural communication like sign language, or some other mechanism. However, there is likely the presupposition underlying such games that some kind of symbolic communication is still possible.

Communication Games are, of course, much more permissive in general of physical and representational differences than Imitation Games, as we test for perceived communicative ability, instead of exact imitation of human attributes. Note that neither the Imitation Game nor the described Communication Test say anything about whether or not an entity under examination has ‘intelligence’ if it fails the test. As mentioned above, false negatives are possible in both tests. Such issues appear to be inherent in both Imitation Games and in such Communication Tests.

Note that one could imagine a Communication Game, where a non-human species is enabled to participate in the test by the use of specially designed equipment (such as a lexigram board), suitable to their particular embodiment. However, I make no particular claims as to the suitability of such tests, only that attempting to conduct them seems possible, at least in theory.

It is relevant to note that some researchers have argued that it is our use of language and symbols that separates us from other species (Deacon 1997). With this in mind, it may turn out that an “anthropocentric flaw” for any Communication

Game exists in the context of attempting to communicate with entities (artifactual or otherwise) that are very different to us: The use of language or symbolic communication itself. However, this issue can be avoided to some extent by simply restricting the objective of test as a partial search for human-like intelligence, rather than any kind of intelligence. Note that such arguments do not really factor into any discussion of the Turing Test as it seems to be unintentionally limited to humans (or near human duplicates) only.

Conclusion

This paper described the notion of a Communication Test, which is an attempt to address the sub-cognitive limitations inherent in the Turing Test and its near relatives. It does this by shifting the goal posts of the Turing Test and its close relatives from imitation to communication, and by requiring this communication to meet an objective measure of success. Some details and examples are explored to examine how this might work. The objective of a Communication Test is not to convince an Examiner that an Examinee is a human being, but simply to attempt to achieve a goal via sharing meaning with a conversation partner. If an Examinee passes an equivalent number of such tests to a human being, then the Examinee is considered to have passed the test at a human level of performance. As it is possible to pass the test and be obviously non-human, the philosophical questions of 'thinking', 'consciousness', and so forth that are purportedly examined by the original Turing Test are still open for debate and/or relevance in such a test.

The Turing Test is often considered as a possible test for 'intelligence' as well as 'thinking'. However, due to anthropocentric limitations inherent in the test for imitation, it may be considered as essentially a test for "humanness". Once we step away from testing for humans only, or accept identification of a non-human test subject as not being grounds for failure in the test, it seems difficult (if not impossible) to avoid introducing some anthropocentric bias into the test. The Communication Tests described in this paper are an attempt to allow an obviously non-human test subject to be tested for human-like and human-level intelligence, whilst at the same time attempting to minimise our own biases. Like the Imitation Game and near relatives, the use of such a Communication Test also allow us to side step the philosophically slippery slope of defining 'intelligence', 'consciousness', 'thinking', and so on. However, once we allow acceptable identification of a non-human in the test, the philosophical implications of the test become more difficult to discern. In spite of this, such communication tests have an advantage over the Turing Test in that they allow us to avoid some of the key sub-cognitive and anthropocentric limitations, and may have a practical advantage over the Turing Test:

Attempted instantiations of the Turing Test, such as the Loebner Prize, have received criticism in the past, in part because of the entrants' focus on 'fooling' the Examiner using relatively low-tech solutions, rather than making serious attempts towards demonstrating intelligent behaviour. Working towards passing an instantiation of the described Communication Test might allow a similar test to be conducted, but with a focus more directed towards ‘intelligent’ solutions. This style of test might find an initial affinity with some researchers interested in AI and related areas, particularly those focused on human-style communication and human-machine interaction. The exact form of such test instantiations, and their degree of philosophical relevance, seems a suitable topic for future research and debate.

Acknowledgements The author would like to thank Alan Blair and David Chalmers for their words of encouragement regarding an early draft of this paper. The author would also like to thank the anonymous reviewers whose feedback he found helpful.

References

Deacon, T. (1997). *The symbolic species*. New York: W. W. Norton and Company.

Dick, P. K. (1968). *Do androids dream of electric sheep?* Garden City, NY: Doubleday.

Fisher, R., Patton, B., & Ury, W. (1992). *Getting to yes: Negotiating agreement without giving in* (2nd ed.). Boston, MA: Houghton Mifflin.

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

French, R. (2000). The Turing Test: The first fifty years. *Trends in Cognitive Sciences, 4*(3), 115–121.

Harnad, S. (2001). Minds, machines and Turing: The indistinguishability of indistinguishables. *Journal of Logic, Language, and Information* (special issue on “*Alan Turing and Artificial Intelligence*”).

Lakoff, G., & Johnson, M. (2003). *Metaphors we live by* (2nd ed.). Chicago: University of Chicago Press.

Loebner, H. (1994). In response [to Shieber: Lessons from a restricted Turing Test]. Accessed April 19, 2009, from http://www.loebner.net/Prizef/In-response.html

Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A modern approach* (2nd ed.). NJ: Prentice-Hall, Inc.

Schank, R. (1987). What is AI anyway? *AI Magazine, 8*(4), 59–65.

Shieber, S. (1993). Lessons from a restricted Turing Test. *Communications of the Association for Computing Machinery,37*(6), 70–78.

Turing, A. (1948). Intelligent machinery. *Machine Intelligence, 5*, 3–23. (Manuscript written in 1948, published in 1969)

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.