Dad jokes, D.A.D. jokes, and the GHoST test for artificial consciousness

Steven Gimbel[*]
Clifton Presser[†]
Paul Mogianesi[‡]

Abstract
The ability of a computer to have a sense of humor, that is, to generate authentically funny jokes, has been taken by some theorists to be a sufficient condition for artificial consciousness. Creativity, the argument goes, is indicative of consciousness and the ability to be funny indicates creativity. While this line fails to offer a legitimate test for artificial consciousness, it does point in a possibly correct direction. There is a relation between consciousness and humor, but it relies on a different sense of "sense of humor," that is, it requires the getting of jokes, not the generating of jokes. The question, then, becomes how to tell when an artificial system enjoys a joke. We propose a mechanism, the GHoST test, which may be useful for such a task and can begin to establish whether a system possesses artificial consciousness.

Keywords: artificial intelligence, humor, consciousness, Douglas Hofstadter, Turing test[§]

[*] Department of Philosophy, Gettysburg College (Gettysburg, Pennsylvania, USA); sgimbel@gettysburg.edu (corresponding author)

[†] Department of Computer Science, Gettysburg College (Gettysburg, Pennsylvania, USA); cpresser@gettysburg.edu

[‡] Department of Computer Science, Gettysburg College (Gettysburg, Pennsylvania, USA); mogaipa01@gettysburg.edu

[§] Received on May 20th, 2021. Accepted on June 24th, 2021. Published on June 30th, 2021. doi: 10.23756/sp.v9i1.612. ISSN: 2282-7757. eISSN: 2282-7765. ©S.Gimbel, C.Presser, and P.Mogianesi. This paper is published under the CC-BY license agreement.

1. Introduction

Donald Mitchie (1993) argues that we need to recast the traditional distinction between the easy and hard problems of consciousness when considering artificial intelligence and instead think in terms of "the problem of artificial intelligence" and "the problem of artificial consciousness." The former, he argues, requires a successful Turing test; while the latter, he contends, demands something different, what he terms a successful "Searle test."

In other words, we can think of the question of artificial intelligence as being comprised of four different questions:

(1a) Can a constructed artificial system be intelligent?
(1b) What test would separate the intelligent from non-intelligent constructed systems? (2a) Can a constructed artificial system be conscious?
(2b) What test would separate the conscious from the unconscious constructed systems?

Questions 1a and 2a belong to the computer scientists, whereas questions 1b and 2b belong to the philosophers. The questions 1b and 2b ask how we define the terms, where we draw the line, and how can we operationalize it. Questions 1a and 2a are challenges to develop systems that cross the line or theoretical arguments providing reasons to believe the line can never be crossed. Indeed, the answers to 1b and 2b in no way presume that the test may ever be passed. Even if one could show, as some argue, that machine consciousness is impossible, the claim that the answer that 2a is "necessarily no" presumes that there is an answer to 2b (an in-principle Searle test, in Mitchie's terminology) which we can demonstrate that no computer could ever successfully complete. To contend that 2a is *a priori* false requires accounting for the necessary existence of a gap between the upper limit of technological systems and what would be required by a hypothetical system that could pass the proper Searle test.

Today, with machine learning as an established subfield of computer science, we can be assured that question 1a, the problem of artificial intelligence, has been answered in the affirmative. This is not to say that there is not interesting philosophical work still to be done around question 1b. The line has surely been crossed, but the questions "where exactly was the line" and "why put it there?" remain interesting.

Mitchie refers to an answer to 1b as a "Turing test," but that is in the general sense of that ambiguous term. Whatever the correct answer to 1b turns out to be, it will remain an issue for Turing scholars to assess how Alan Turing's various versions of the imitation game, and the assorted extensions of it, turn out to relate to that line. The questions 2a and 2b, on the other hand, both remain tantalizingly open. To 2a, we have to answer "no." But it is unclear whether by "no" we should mean "not yet" (the empiricist position) or "not possible" (the a priori position). To distinguish between these, we would need to assess the possibility that one could in principle construct a system that would pass Mitchie's Searle test. But that claim requires having the proper Searle test, in other words, the answer to 2a presupposes an answer to 2b.

There are two approaches to 2b, that is, two very different visions of what a successful Searle test would look like. On the one hand, there are the "behaviorists" who argue that a successful Searle test would be of the same sort as the successful Turing test, only include more intricate criteria. On this view, one can infer consciousness from something like an imitation game only including, say, output that requires some capacity beyond learning, e.g., creativity. On the other hand, there are the "structuralists," like Pentti Haikonen who contend that if machine consciousness is possible, tests for it would have to be concerned with internal structure of the system and not inferred from output of it. We can infer nothing from the results, only from a similarity of structure.

A lot, therefore, hangs on the formulation of a successful Searle test. We explore a novel approach, the GHoST test, which does not offer a complete answer to 2b, but does allow us to establish the intellectual neighborhood in which the successful Searle test would have to reside, through something akin to the intermediate value theorem, that is, we will be able to see how passing the GHoST test in one sense is insufficient for consciousness and how passing it in a different way would be clearly sufficient. Hence, the proper Searle test, whatever it turns out to be, will have to draw the proper line somewhere in between.

The GHoST test is a behaviorist test, which differs in important ways from the standard approach (and which will require an additional structural element). A strand of the behaviorist research program to develop the proper Searle test focuses on computer creativity, with a sub- strand devoted to computer humor creation – to be truly funny, they contend, requires consciousness. We argue that this approach provides too weak a criterion to act as a properSearle test. Instead of looking for a system capable of generating laughs in humans, we should instead look for a system capable of creating digital auto-didactic, or D.A.D. jokes, that is computer- generated dad jokes. The emergence of D.A.D. jokes, jokes known not to be funny but told anyway, may, in fact, constitute evidence for machine consciousness. We contend that this approach at least succeeds in avoiding the standard problems and accounting for models of consciousness that are different from human consciousness.

2. Empiricist and a priori arguments concerning machine consciousness and the need for a Searle test

Not only is the question concerning computer consciousness open, but the question as to whether it is even a question remains open. Some, like Douglas Hofstadter contend that it is an empirical question. We can conceive of systems, at least hypothetically, that would, pass any successful Searle test, so it is a possibility. On the other hand, there are those like Louis Marinoff who argue that given any reasonable definition of consciousness, any machine will necessarily fall short.

The classical argument against computer consciousness comes from Ada Lovelace. "The Analytical Engine has no pretentions to *originate* anything. It can do *whatever we know how to order it* to perform (Quoted in Turing, 1950, 450 – italics in the original)." A necessary condition for consciousness, Lovelace contends, is creativity. Hofstadter, who wrote his (2009) essay "Essay in the Style of Douglas Hofstadter" in the style of Douglas Hofstadter, argues that the ruling out of the meaningfulness of a creativity-based Searle test on *a priori* grounds is illegitimate. He considers David Cope's EMI which produces novel compositions in the style of human composers whose work is entered in as input. Hofstadter then posits a hypothetical, maximally successful version of the system which created new pieces of music such that the best-trained experts could not differentiate from pieces by great composers, say, Bach or Mozart. He then further imagines an analogous system which would be similarly successful in the development of scientific results. We can imagine a machine capable of writing papers with new discoveries in physics whose creative and mathematical approach is so much in the style of Albert Einstein that even physicists and Einstein scholars could not tell if it was Einstein or the machine who produced it.

Using this thought-experiment, Hofstadter proposes a *reductio ad absurdum* argument designed to bolster the intuition that the limit case of a creativity-based Searle test should be considered a successful answer to the Lovelace objection. If we had the Einstein imposter of a computer, then surely this version of the imitation game would be successful in meeting Lovelace's concern. This type of creativity is so impressive that anything capable of it surely must have a mind. Using discoveries in physics of the level and in the style of Albert Einstein as a limiting case, Hofstadter concludes that it is at least in principle possible – extrapolating from technology we have today – to envision a machine we would be forced to conclude has sufficient creativity that demands we accept it as conscious. By positing such discoveries, we can now imagine possible creative output from a machine that, if it were to be observed, would surely satisfy even the most hardened Lovelacian critic.

On the other hand, there are those like Marinoff who argue that such imaginings are irrelevant fantasies. Machines are machines and we have provable theorems concerning computability and those produce upper-limits that will always and necessarily fall on short-side of the line constructed by any successful Searle test.

> When it comes to performing quantitative tasks in competition with humans including playing games such as checkers and backgammon, or even chess and Go, the computer is no longer the underdog, but the overdog; not yet and perhaps never to be a Nietzschean *übermensch* in evolutionary terms, but demonstrably an *überhund* at parlour games (74).

Computers may have abilities that we consider cognitive and may be superior to humans in the rate and accuracy of such computational procedures, but there is a necessary difference between the organic and constructed mind that necessarily keeps computers on the weak side of the Searle test.

> I submit that a–perhaps *the*–salient difference between computer versus human performance lies not merely in *what* they can and cannot do, but rather in *how* they attempt to do what they can and cannot do. In methodological terms, the computer is an entity that strictly follows instructions, while the human is a being that constitutionally disregards them. Computers do exactly and only what they have been instructed to do, whereas humans are capable of an inexactitude that includes, but is not restricted to the self- prompted or unconscious misinterpretation, omission, permutation, and modification of members of a given instruction set (ibid.)

There is, in this quotation, a vague appeal to an intuitive Searle test which, Marinoff argues, will necessarily be failed by any computer no matter how powerful the hardware and subtle and clever the software.

He considers two competing arguments based upon Church's theorem. The first is the empiricist argument which he frames as:

1. All and only intuitively computable functions are Turing computable (Church's theorem).
2. Understanding and meaning are intuitively computable functions we just haven't figured out how to compute them yet. (Empiricist belief)

Therefore, understanding and meaning are Turing computable. (Strong AI)

Contrast this with the a priori argument:

1. All and only intuitively computable functions are Turing computable (Church's theorem).

3. Understanding and meaning cannot be intuitively computable functions. (a priori belief)

Therefore, understanding and meaning are not Turing computable. (Denial of strong AI)

We do not have a proof of Church's theorem, but as both views depend upon it equally, let us assume it. The question is whether we have good reason to believe 2. or 3., and Marinoff argues that "we have reason for supposing the understanding and meaning are not intuitively computable," what he terms the "reverse Turing test, furnishes one such reason (76)."

> Suppose that a human (H1) is given a set of instructions (S1) which, if faithfully executed, would result in the imitation of a Turing machine (T1). But suppose that the human makes meaningful mistakes in their execution. Now, we ask whether we can build another Turing machine, T2, such that T2 can make meaningful mistakes (ibid.).

The answer must be yes or no. If no, then no strong AI. If yes, then T2 must have been given a set of instructions S2 which it faithfully executed, thereby not truly having committed meaningful mistakes, but rather have made no mistake in imitating humans who make meaningful mistakes. He has revived the Lovelace objection.

Two points need to be taken away from this. First, both Hofstadter and Marinoff presuppose, a line but do not establish a Searle test telling us where it is. Hofstadter gives us a thought experiment whose conclusion is presumed to have passed any reasonable line, where Marinoff gives us an argument that presumes that any mere instruction executing system must not have crossed any reasonable line. For either of these arguments to be complete, the line needs to be established. Second, while the two disagree on whether crossing the line is possible, they both provide an interesting insight that may allow us to think more clearly about where the line is. Marinoff focuses on human's proclivity for making meaningful mistakes. Hofstadter has made a similar claim, that we will know that we have artificial consciousness when we find humans and computers making the same sort of mistakes. They may both be correct that the missing element in the conversation, the successful Searle test could be something related to a reverse Turing test.

3. Possible forms of a Searle test

There are two different approaches to testing for artificial consciousness, behaviorist and structuralist. The behaviorist tests are those that draw the line based on the output of a system. Such tests are appropriate for Turing tests (in Mitchie's sense) because, unlike in the case of consciousness, intelligence does not fall prey to the Lovelace objection. To act intelligently is to be intelligent. If, for example, we were to take learning as a sufficient condition for intelligence, one cannot imitate learning without learning. In this way, intelligence is like singing. The only way to imitate singing is to sing. As such, the question is which sorts of cognitive behavior are the correct ones with which to draw the line and do we have examples of artificial systems engaging in it.

But consciousness is a completely different matter. Where intelligence is a behaviorist notion and thereby open to behaviorist testing; consciousness seems to require something not directly observable, something within the system. One could create an object that imitates being conscious without being conscious, examples are plentiful from Eliza and Siri to *Weekend at Bernie's* and Munch's Make-Believe Band at Chuck E. Cheese.

Behaviorists argue that the problem of consciousness with respect to artificial systems is no different than the problem of other minds with regard to seemingly fellow humans. Going back at least as far as René Descartes, the inverse problem of consciousness has been asked: how do I know that the people I believe to be conscious are not just automata? We were at that time, of course, restricted in this matter to behavioral data. B. Jack Copeland (2003) points out that Descartes' protégé Géraud de Cordemoy, in his book *A Philosophical Discourse Concerning Speech*, used the Cartesian insight to anticipate Turing's imitation game*:*

> To speak is not to repeat the same words, which have struck the ear, but to utter others to their purpose and suitable to them. …[N]one of the bodies that make echoes do think, thought I hear them repeat my words...I should by the same reason judge that parrots do not think neither….But not to examine any further, how it is with parrots, and so many other bodies, whose figure is very different from mine, I shall continue the inquiry...I think I may...establish for a Principle that...if I find by all experiments I am capable to make, that they use speech as I do,...I have infallible reason to believe that they have a soul as I do (quoted in Copland, 10).

At this time, the concept of mind and soul were considered identical, so de Corduroy has produced a principle which he deemed sufficient for determining if something has a mind: "If a non-human thing has the capacity to use speech as humans do, a condition subject to experiment, then that thing possesses a mind." When they make unpredicted conversational contributions that make sense to us and are human-like in their content, then we have reason to infer that we are interacting with a second, distinct intelligence.

The behaviorists contend that if this test is good enough for the opacity of the human mind (for those other than yourself), then it should be good enough for the general case. Of course, the behavior could not merely be conversational imitation, but would need to be something much more cognitively intricate to give sufficient evidence of consciousness. A standard criterion for the extended de Courtemoy approach to a Searle test involves computer creativity. Truly creative output, they argue, requires a mind and so if we can find the right sort of creative endeavor, we could formulate a creativity-based behaviorist Searle Test.

This is what Hofstadter is arguing in (2009). If a computer could produce multiple papers with legitimate scientific discoveries based on reasoning that experts could not tell from that of Albert Einstein, then this version of the imitation game would be successful in meeting the critics' concern. Discoveries in physics of the level and in the style of Albert Einstein are a limiting case. By positing such discoveries, we can now imagine possible creative output from a machine that, if it were to be observed, would surely satisfy even the most hardened Lovelacian critic.

Others have trod Hofstadter's behaviorist path by returning to Cope's musical approach, but the type of musical output in the hope that we would not need a limiting case at the level of Einsteinian physics to undermine the apriority of the Lovelace objection. Antonio Chella and Riccardo Manzotti (2012) contend that we do not need breakthroughs in physics, a sufficient demonstration of computer creativity would be found in a computer that could be an integral part of a jazz ensemble. It is one thing, as Cope's EMI system does, to compose scores, but to successfully improvise musically with a band, that is, to interact in real time with humans who are swinging and contribute artistically in a fashion that would be non-differentiable from a human musician, that should be enough to satisfy the critics and count as computer creativity.

Still others have changed the artistic medium altogether. As Turing sets out, critics contend that among the things a machine could never do is, "Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as a man, do something really new (1950, 447)." The one that seems most tractable is "have a sense of humor." Researchers including Huma Shah and Kevin Warwick (2017), Graeme Ritchie (2015), and Margaret Boden (2009) have therefore been focusing on using humor generation as the creativity- basis for an empirical undermining of the Lovelace objection. If we can create a computer with a sense of humor, then that sort of creativity should lead us to posit consciousness.

Ritchie (2009) notes that the operative phrase "sense of humor" is itself not rigorously defined and as such determining when a computer can be said to have a sense of humor is thus also underdetermined. To use this property in an argument for strong AI, it needs to be rigorously testable. While the notion may be vague, intuitively it seems satisfied by a person who generates original humorous utterances that others find funny. As such, if computers can create new humor that is genuinely funny, then we may have an empirical justification that undermines the Lovelace objection.

How close are we? Shah and Warwick (2017) review a number of recent attempts. Their conclusion is that some of the best computer-generated humorous utterances are somewhat amusing. In other words, we certainly do not have the empirical evidence that would be needed to counter the claim of Lovelace critics, but what they argue we may have is supporting evidence that the research program is moving in a positive direction. In other words, we have some evidence that someday we might have the evidence we need to conclude a positive creativity-based behaviorist Searle test.

Those who take the Lovelace objection to be conclusive contend that no behaviorist test will ever be a successful Searle test. Because consciousness requires internal qualities of an individual (e.g., internal monologue, sense of self, volition) and because these are always opaque, it is always possible to imitate them with truly possessing them. The only system that we know for sure possesses it is the human brain. From a physiological system comprised of mere atoms, consciousness arises as an emergent property. As such, the line goes, the only possible way to be assured that artificial consciousness exists would be to have a system that intricately and accurately models the structure of the human brain.

On this view, the only sort of test that is possible would be structural. Neuroscientists work on their end to develop the "neural correlates of consciousness," that is, the minimal set of interrelated structural elements in the brain that allow consciousness to emerge (Crick and Koch 1990). Computer scientists then work to build an isomorphic artificial neural network. If we know what the minimal structure is for a mind to emerge, then we have our lower limit on what we need to model, go and do it.

A different version of the structuralist approach is the "cognitive approach" taken by Pennti Haikonen who argues that the isomorphism does not have to map component onto component, but rather function onto function.

> Conscious robot cognition calls for information integration and sensorimotor integration and these lead to the requirement of an architecture, the assembly of cross-connected perception/response and motor modules (Haikonen 2012, 4).

Haikonen's argument is that what gives rise to consciousness is to be found in the functional modalities of the mind. If we can construct components that do what the parts of the brain do, and have them do them interactively the way an organic human brain does, that should be sufficient to give rise to the emergent property of consciousness even if we cannot map neurons onto artificial neurons. In this way, he has developed a computerized cognitive system with internal monologue, which seems prima facie to be a property of conscious entities. Haikonen sees this achievement as a step in the right direction, but insufficient.

> How about machine consciousness? Self-consciousness is not yet emulated here, as the simulation system does not have episodic memory for personal history nor body reference for self-concept ("I") and therefore is not able to perceive itself as the executing agent. Even though the system has the flow of inner speech and inner imagery and it operates with them, it is not yet able to report having them. It is not able to produce much towards the response "I have inner imagery" or the consequence "I think – therefore I exist". Obviously, this kind of a report would only count as a proof of self-consciousness if it can be seen that the system is producing it meaningfully, i.e. the system would have to be able to perceive its inner imagery as such and it would have to possess the concepts like "I", "to have" and "inner imagery". The mere reproduction of preprogrammed strings like "I have inner imagery" would not count as a proof here (Haikonen 2000, 8).

Internal monologue may be necessary, but is not sufficient for consciousness.

However, Haikonen does point out an issue with both the behaviorist and structuralist approaches to a Searle test – neither seems sufficient in and of themself. If one has a system that appears to pass a behaviorist Searle test, one always has to worry about the Lovelace objection, that is, was the machine simply programmed to produce that seeming sign of consciousness. Is it a mere imitation? Similarly, with a structuralist Searle test, just because you have a system that structurally resembles an organic system in which consciousness emerges, do you, in fact, have consciousness? It seems that any system that passes a structural Searle test would still have to demonstrate its consciousness in some behavioral fashion. Hence, a successful Searle test seems to need both a structuralist and a behaviorist component.

4. The GHoST Test

Those who try to use humor generation as a behaviorist Searle test contend that the creativity necessary to create truly funny jokes should be seen as an answer to the Lovelace objection, that is, the ability to repeatedly construct novel comic utterances that are genuinely funny is to have a sense of humor. Only conscious beings can have a sense of humor. Therefore, this works as a Searle test.

The problem, of course, is that it does not answer the Lovelace objection. It certainly seems conceivable that one could create an algorithm that would analyze a category of joke, identify those properties that funnier jokes share, and use that to form new versions. Consider the old chestnut, “I saw a man in the park with a telescope.” Humans are quick to see one ambiguity in this sentence, that you spied a man carrying a telescope versus that you spied the man through a telescope. Artificial semantic evaluation differs from human analysis in just as quickly picking up on the peculiar interpretation whereby you used the telescope to saw the man in half. Human consciousness is thereby a hindrance to finding certain linguistic interpretations which artificial means may find easily due to the lack of influence of certain sorts of psychological priming.

It does not seem absurd that such ambiguities might be used by a computer-based joke generator as the basis for pun-based jokes which, because of the learned effective joke structure, would be as funny as those of a human with a developed sense of humor, but because of the human’s disinclination to see these particular ambiguities would thereby be novel jokes. So, we would have a generator of funny new jokes, but because it is all done algorithmically, we surely do not have confidence of the system’s consciousness as a result. Having a sense of humor, thereby, does not seem to be a legitimate Searle test.

But the notion of a “sense of humor” is ambiguous. Indeed, Martin and Ford (2018) distinguished between three notions of that phrase. When we say someone has a sense of humor we could mean (1) that the person possesses an active faculty of humor appreciation, that is, that they like a good joke, (2) that the person is skilled at humor delivery, that is, they are the life of the party and know how to keep a crowd laughing, or (3) that the person has an active faculty of humor production, that is, that the person is a generator of novel comic acts. These are three very different properties, yet in common parlance all of them are described by having a “sense of humor.”

The one that is relevant to a possible Searle test is not the first, ability to create jokes, but rather the third, the ability to get jokes. We all have jokes that we are thoroughly embarrassed to find funny. We know these jokes are poorly constructed, morally problematic, juvenile, or just plain stupid; yet, we cannot stop ourselves from snickering at them in spite of ourselves. It is that experience of finding the funny that is the real mark of intelligence, whether or not we are capable of creating the funny or bringing the funny. What we are looking for with the Searle test, in essence, is Gilbert Ryle’s “ghost in the machine,” but we are looking in the wrong place with computer joke generators because while we might be able to make the machine crank out jokes, it would only be the ghost that would find them funny.

Unfortunately, we run into the Lovelace objection. We could certainly design a system that recognizes jokes and that learns what jokes humans tend to enjoy in jokes and then which mimics human laughter at an appropriate level: a guffaw for successful slapstick, a haughty chuckle at a clever witticism, and a disapproving eyebrow raise for a tight pun. This, again, would be mere imitation that would never satisfy the Lovelace critic – nor should it. We need a different approach for a comedic hunting of the ghost.

That different approach will pull together insights from several disparate sources and bring them together in a proposal for a new sort of test. The first insight came from Douglas Hofstadter (and is echoed in Marinoff). When our institution was fortunate to host Hofstadter for a visit a few years back, he responded to a question about artificial intelligence with a statement to the effect that we will know we have strong artificial intelligence, not when we see computers doing the sorts of things we can do, but when we see them making the sorts of mistakes we make. Humans and computers both make mistakes, he pointed out, but they make radically different sorts of mistakes. The output when there is a bug in a program looks nothing like the sorts of cognitive mix-ups we see in humans.

Consider the catalogue of error-types that Hostadter and David Moser (1989) collected. These include categories like malapropisms: “I like a magazine with good, objectionable reporting,” spoonerisms: “tea and flick spray,” infelicitous metaphors: “Welcome to Israel, a Mecca for tourists,” and metaphors: “That was a breath of relief.” They have a range of categories and loads of examples. Part of what makes these instances of human error amusing is that we recognize them in our own experience. They are, indeed, human. The brain works in a specific way and is prone to certain sorts of mistakes based upon that wiring.

As such, we can learn about the wiring through the sorts of errors to which it is prone. Computers are wired differently and so they make different sorts of errors, errors that we tend not to see in people. When we can construct machines that make errors more similar to ours, Hofstadter's line went during his response at the talk, then we will have machines with wiring more like ours and that is when we can lay claim to having developed artificial consciousness.

The GHoST test will appropriate the insight that we need to examine computer failings instead of successes for signs of intelligence, but change the sort of errors examined. Hofstadter's interest is in cognitive-linguistic miscues. We will argue that deeper ramifications are to be found in a quite different sort of mistake. The clue for that is to be found back in Turing (1950). In his consideration of "the argument from informality of behavior," he discusses the lack of rule- boundedness that we find in lived human choice-making.

> One might for instance have a rule that one is to stop when one sees a red traffic light, and go when one sees a green one, but what if by some fault both appear together? One may perhaps decide that it is safest to stop. But some further difficulty down the road may well arise from this decision later (Turing 1950, 452).

Turing's point is that the number of potential situations in which one may find oneself is potentially unlimited and so we would potentially require an unlimited number of behavioral rules. But we cannot know a potentially unlimited number of rules, meaning that human behavior is not rule-bound.

Perhaps this is true, perhaps it is not. The particular objection is not our concern. Rather, it is this shift from cognitive processing to social behavioral rules that is important for this matter and Turing's invocation of them provides the spark. But while all human behaviors may not be rule- based, human conversational behaviors may be. This is the well-known view championed by H.P. Grice (1975). Conversation is a cooperative endeavor and as such requires rules to function properly. Grice sets out four maxims: Maxim of Quantity: "Make your contribution as informative as is required (for the current purposes of the exchange)," Maxim of Quality: "Try to make your contribution one that is true," Maxim of relation, "Be relevant," and Maxim of Manner: "Be perspicuous." To be a responsible conversant, one should follow the rules as obeying the maxims will allow the conversation to be maximally successful and efficient.

However, these rules are sometimes broken in the course of conversation. When someone observes such an infelicity on the part of their co-conversant, there is a decision to be made. One could consider one's conversational partner to have violated the Cooperative Principle because the person is uncaring, uncouth, or insufficiently well-versed in proper conversation. But, one might reject this belief and consider one's co-conversant to be fully dedicated to the cooperative conversational project. If this were true, such an infelicity would be sure to be noticed as an unexpected act. Perhaps the rule was violated as some sort of signal that the conversational context is about to change radically and this requires a change of behavior on both of our parts. Perhaps, the colleague whose affairs we had been discussing is coming to join the conversation necessitating a rapid change in topic. The violation of a conversational maxim may be an error that has a cooperative function. To discern the goal of the intentional error and thereby understand the intent of the speaker requires an inference, a conversational implicature, on the part of the listener.

So, from Hofstadter, we take the insight that errors may be more informative than successes. From Turing, we get the shift from cognitive-linguistic blunders to behavioral ones. From Grice, we take the notion that there may be inferences to be made from willful, intentional violations of cooperative conversational rules. What is left is an appropriate rule for comic conversation. For this, we turn to Peter Singer. Despite nearly two and a half centuries of devastating counter-examples to the principle of utility utilizing examples ranging from slavery to infanticide to bestiality, Singer continues to doggedly maintain complete devotion to it. One cannot but respect such fidelity. We can use it to formulate a Gricean sort of maxim for comic conversation that demands the maximization of comic utility, “only tell jokes that you have good reason to believe your audience will find amusing.” Joke only so that you maximize overall utility.

The advantage of this sort of rule is that it can be built into a machine learning algorithm for a computer. If we design a computerized joke creation program, we can have people rate the output in such a fashion that the program will learn what people do and do not find funny. In this way there should be an observable learning curve according to which the machine gets progressively funnier. We should see a stable upward trend in the funniness of the computer’s output. This is not to say that there will not be misses. There will certainly be instances that are below the trend line as there will be those which form the sorts of advances from which the machine will learn.

Putting together the insights from **G**rice, **Ho**fstadter, **S**inger, and **T**uring, we can construct the GHoST test wherein willful violation of a utilitarian principle concerning humor in conversation will provide us with legitimate warrant for an inference of true machine intelligence. We can show that this test suffices to attribute intelligence to humans and will, thus, be useful in assessing a specific set of possible artificial products.

5. Dad jokes, bad jokes, D.A.D. jokes and B.A.D. jokes

Humans are capable of violating behavioral rules. We do it all the time...some of us more than others. The stereotypical middle class American father is widely known to violate the Singer-Grice comedic maxim. Corny, clean, often pun-based jokes that the father knows will not be enjoyed by their children (particularly if teenagers and especially in the company of their friends) are known in colloquial terms as “dad jokes.” Dads fully know the reception that their jokes will receive, and yet tell them anyway. This is a clear violation of the generally accepted rule to only tell jokes that will maximize overall utility, but that does not stop Pop. Why do dads tell these unappreciated dad jokes? Because they want to. Because they find the charming little jokes funny. It is a selfish, albeit harmless, expression of volition. Dad is amusing himself. Dad jokes are bad jokes, but dad doesn't care.

Dad is a biological entity (no matter what mom sometimes says). In telling jokes that he knows others will not enjoy, but which amuse him, he is acting autodidactically. Bad jokes told by organic, living beings are "biological autodidactic" jokes, or B.A.D. jokes. Dad jokes are not only bad jokes, they are also B.A.D. jokes. B.A.D. jokes being intentional acts that violate a behavior rule are following de Courdemoy's condition, evidence that dad is intelligent (again, no matter what mom sometimes says). It is a specific sort of verbal act that requires not only an active sense of humor (in the first sense) but also volition. Dad jokes combine two elements that seem individually indicative of intelligence. Combined, they surely are even more so.

If a computer created a joke that it knew was below the standard of humor appreciated by humans, but which it decided to tell anyway, we would have an example of a digital autodidactic joke, or D.A.D. joke. D.A.D. jokes, like dad jokes, are bad jokes; but D.A.D. jokes, unlike dad jokes, are not B.A.D. jokes.

If we can use dad jokes in combination with an extended de Cordemoy condition to infer intelligence in dad, then the same ought to hold for D.A.D. jokes and their non-organic originator. If a computer acts autodidactically, then it acts with a will and only thinking things have a will. If that artificial mind possesses the desire to tell a joke it knows the audience will not enjoy, then we have reason to believe that it was told because the program itself though it was funny. This would be evidence that we are dealing with something with a sense of humor (in the first sense). D.A.D. jokes would be evidence in favor of artificial consciousness.

How then do we know when we have a D.A.D. joke? We can establish a lower bound employing a joke-generator that we know does not pass the structural element of a Searle test for consciousness (the sort discussed in Shah and Warwick 2017), but is capable of learning based upon human reactions to the jokes it constructs. In this way, there should be a positive curve with the jokes becoming funnier over time, although certainly there ought to be expected clunkers in the bunch.

Syntactic and semantic evaluation of successful jokes would produce generalized joke structures which would create templates for further new jokes. One would have to expect "I saw the man in the park with the telescope" sorts of instances wherein the novel generated joke conforms to the structure according to which it ought to be a joke that humans find funny, but which, because of our cognitive make-up, humans tend not to find funny. For example, suppose a system developed such a template for successful jokes and used it to iteratively create a joke "fractal," that is, a joke in which embedded versions of the joke structure, taken together form an example of the structure itself. Such a joke might be capable of analysis that renders it technically successful, but beyond the capacity of the human mind to process sufficiency to garner a reaction. Or perhaps, a system could develop a combination of templates, again whose complexity undermines its comedic success despite being consistent with the a posteriori results working as the foundational input into the system. The presentation of such instances would be an example of intelligence without consciousness, yet an expectation of jokes being funnier than they are.

In other words, we would expect the artificial joke generator to "think" that these jokes are funnier than they, in fact, are. Because the system, by supposition, does not pass the structural side of the Searle test, we know that the computer, despite having good reason to "believe" the joke is funny does not find the joke funny. This is thereby the lower bound on the behavioral side of the Searle test.

Suppose, on the other hand, for the sake of argument, that we have a system that does satisfy the structural element of a Searle test. Now, we are looking for a behavior that indicates the sort of cognitive properties one associates with consciousness, the "I think, therefore I am" moment. If such a system included the same sort of joke generator that we posit in the prior example, then, again, we ought to expect a learning curve, with a positive humorousness trajectory with the occasional clunker. Clunkers would not be indications of D.A.D. jokes since we would also see them in the prior case.

What one would need to see is repeated generation of a related set of similar jokes that conform to the rules of quality joke generation, but are of the "I saw the man in the park with the telescope" variety. The joke generator would have learned through syntactic and semantic evaluation how to create jokes. Through the response to the jokes, it would have learned that there is a gap between successful jokes of the format and "I saw the man in the park with the telescope" examples, that is, there is a gap between what ought to be funny and what is funny. To continue to explore the "ought" line and not surrender it to the "is" line, despite negative conditioning, would be the digital equivalent of making dad jokes, that is, they would be D.A.D. jokes. The distinction in semantic processing should, on the condition that we do hypothetically have an instance of artificial consciousness, give rise to a different sense of humor in the first sense. The system wouldbelieve that jokes humans do not fund funny, are, in fact, funny. It would find them funny and believe that we have the defective sense of humor. Humans are cognitively limited, just not smart enough, to understand how funny these jokes, in fact, are.

The claim is not that there is or ever will be such a system. Again, what we are looking for here are the conditions for a Searle test. One can hold the a priori position that there could never be artificial consciousness, but in doing so one still requires a successful Searle test to say that computers could never get to that line. The second case is an upper limit on the behavioral element of such a test. In other words, if there was a system that passed some successful version of the structural element of a Searle test, the passing of the GHoST test as described above would then give us reason to think we have artificial consciousness. The result then, is not that we have sketched a successful Searle test, but that with the GHoST test, we now have a lower and upper bound, between which the behavioral line must be drawn.

References

[1] Boden, Margaret. (2009) "Computer Models of Creativity," *AI Magazine*. Vol. 390, No. 3, 28- 34.

[2] Chella, Antonio and Riccardo Manzotti. (2012). "Jazz and Machine Consciousness: Towards a New Turing Test," *Proceedings of AISB/IACAP 2012 Symposium "Revisiting Turing and his Test: Comprehensiveness, Qualia, and the Real World."* 49-53.

[3] Copeland, B. Jack. (2003). "The Turing Test." *The Turing Test: The Elusive Standard of Artificial Intelligence.* James Moor (ed.). Dordrecht: Kluwer.

[4] Crick, Francis and Kristof Koch. (1990). "Toward a Neurobiological Theory of Consciousness." *Seminars in the Neurosciences.* 2. 263-275.

[5] Descartes, René. (1637). *Discourse on Method.* Indianapolis: Hackett, 1987.

[6] Grice, H. Paul (1975). "Logic and Conversation." Cole, P.; Morgan, J. (eds.). *Syntax and Semantics. Vol. 3: Speech Acts.* New York: Academic Press.

[7] Haikonen, Pentti. (2012). *Consciousness and Robot Sentience.* Singapore: World Scientific.

[8] Haikonen, Pennti. (2000). "An Artificial Mind via Cognitive Modular Neural Architecture."
https://www.cs.bham.ac.uk/research/projects/cogaff/dam00/papers/haikonen.pdf. Accessed 5/9/2021.

[9] Hofstadter, Douglas. (2009). "Essay in the Style of Douglas Hofstadter," *AI Magazine*. Vol 30. No. 3. 82-88.

[10] Hofstadter, Douglas and David Moser. (1989). "To Err is Human; To Study Error-Making is Cognitive Science." *Michigan Quarterly Review.* Vol. 28. No. 2. 185-215.

[11] Jurafsky, Daniel and James Martin. (2008). Speech and Language Processing. Englewood Cliffs: Prentice-Hall.

[12] Marinoff, Louis. (1997). "The Quest for Meaning." *Mind versus Computer: Were Dreyfus and Winograd Right?* Matjaz Gams, Marcin Paprzycki, Xindong Wu (eds.). Amsterdam: IOS Press.

[13] Martin, Rod and Thomas Ford. (2018). *The Psychology of Humor: An Integrated Approach.*
London, Academic Press.

[14] Mitchie, Donald. (1993). “Turing’s Test and Conscious Thought.” *Artificial intelligence*. 60. 1-22.

[15] Ritchie, Graeme. (2009). “Can Computers Create Humor?” *AI Magazine*. 30 (3). 71-81.

[16] Ryle, Gilbert. (1949). *The Concept of Mind.* New York: Barnes and Noble, 1969.

[17] Shah, Huma and Kevin Warwick. (2017). “Machine Humour: Examples from Turing Test Experiments,” *AI & Society*. Vol. 22. 553–561.

[18] Singer, Peter. (1980). *Practical Ethics.* Cambridge: Cambridge University Press.

[19] Turing, Alan. (1950). “Computing Machinery and Intelligence,” *Mind*. Vol. 49. 433-460.