Harnad, Stevan (2011) Minds, Brains and Turing. *Consciousness Online*. February 18th - March 4th 2011 + m4v video + 133 commentaries and responses.

Minds, Brains and Turing

Stevan Harnad
Chaire de recherche du Canada
Institut des sciences cognitives (ISC)
Université du Québec à Montréal
Montréal, Québec, Canada H3C 3P8
&
School of Electronics and Computer Science
University of Southampton
Highfield, Southampton
SO17 1BJ UNITED KINGDOM

The Turing Test. Turing set the agenda for (what would eventually be called) the cognitive sciences. He said, essentially, that cognition *is* as cognition *does* (or, more accurately, as cognition is *capable of doing*): Explain the causal basis of cognitive capacity and you've explained cognition. Test your explanation by designing a machine that can do everything a normal human cognizer can do – and do it so veridically that human cognizers cannot tell its performance apart from a real human cognizer's – and you really cannot ask for anything more.

A machine? Isn't that already a contradiction in terms? Only if you have biassed preconceptions about machines. For "machine" merely means a dynamical system governed by causality. On that score, we too are machines -- for everyone except those who believe that our biology somehow transcends ordinary causality, and that "mind over matter" is somehow an extra, spontaneous force in the universe. (We will not take up this notion till the very end of this essay.)

So we are machines, and Turing simply pointed out that it is therefore our mission to find out the *kind* of machine we are, by explaining how the machine works. His own hunch may have been wrong. He thought that we were mainly computers, and that cognition is computation. So, he thought, the task would simply be to find the right computer program – the one that can pass the "Turing Test" -- able to do anything we can do, indistiguishably from the way we do it.

Searle's Chinese Room. The celebrated thought-experiment in which the philosopher John Searle does all the computations of the computer program that successfully passes the Turing Test [T2] in Chinese demonstrates that cognition cannot "be as cognition does" if the doing consists solely of being able to communicate by email in Chinese indistinguishably from a real Chinese cognizer (even if tested for a lifetime) -- not, at least, if the *means* by which the T2 success is generated is just computation. For computation is implementation-independent: if computation can really do something, it can do that same thing no matter how you implement it physically. And when Searle does the computation that produces the

T2-passing success in Chinese, he is implementing the very same computer program – yet he is not understanding Chinese. So no computer running the same program understands either.

How does Searle *know* he is not understanding Chinese? After all, his is just a thought-experiment. No one yet has actually written a computer program that can pass T2 for a lifetime. Yet we all know that the way to learn Chinese is to learn Chinese, not to learn and execute a computer program that manipulates symbols whose meanings we do not understand, and manipulates them purely on the basis of their shapes, and not their meaning. For that is what computation is, and does. And Searle rightly points out that he (or anyone or anything else implementing the same computer program) would merely be manipulating meaningless symbols under those conditions, not understanding what the Chinese symbols mean, even if they were doing so for a lifetime – despite all appearances to their native Chinese pen-pals.

How is Searle able to make this judgment on the basis of the hypothetical implementation of a non-existent computer program? Let's set aside deeper worries such as whether there could ever be such a successful T2-passing computer program aa well as shallower worries, such as whether, even if there were, Searle could actually do all the symbol manipulations himself. A fundamental question is: How would Searle know whether he was understanding Chinese? If the program was really performing at T2-scale, for a lifetime, then if the lifelong Chinese email interlocutor asked Searle (in Chinese) whether he understood Chinese, the reply (in Chinese) would of course have to be something like: "What an absurd question! Haven't you and I been communicating in Chinese for the past 40 years?"

Yet, when asked, in English, whether he understood Chinese, Searle would reply (quite truthfully) that he couldn't understand a word; he was just faithfully doing the requisite symbol manipulations, according to the rules he had memorized, for the past 40 years. It is the basis for making *that* judgment – that he does not understand Chinese – to which I want to draw attention here, because it really is the heart of the matter. For it calls into question Turing's thesis that "cognition is as cognition does" – at least insofar as language speaking (in this case, writing) and understanding are concerned. Searle is indeed able to do what Chinese speakers/ understanders are able to do, indistinguishably from them. He has all their know-how. Yet he says he is not understanding Chinese: What's missing? *And how does he know it's missing*?

What's missing is what it *feels like* to be able to speak and understand Chinese. And Searle knows, because he is the only one who can know whether or not he can speak and understand Chinese, i.e., know whether or not he has that feeling, regardless of anything else he does or can do. (We will elaborate on this again, below.)

Is this, then, the death-knell for Turing's thesis that "cognition is as cognition does"? So far, this applies only to the very special case of language speaking/ understanding, and only if the means by which the T2 success is accomplished is via computation alone (symbol manipulation).

The Robotic Turing Test. But was the language-only Turing Test, T2, really the one Turing intended (or should have intended)? After all, if the essence of Turing's "cognition is as cognition does" criterion is Turing-indistinguishability from what a human cognizer can do, then *a human cognizer can certainly do a lot more than just produce and understand language*. A human cognizer can do countless other kinds of things in the real world of objects, actions, events, states and traits, and if the T2 candidate could not do all those kinds of things too, then that incapacity would be immediately detectable, and the candidate would fail the test. To be able to do all those things successfully, the T2 candidate would have to be a lot more than just a computer: It would have to be a sensorimotor robot, capable of sensing and acting upon all those objects, etc. -- again Turing-indistinguishably from the way real human cognizers do.

Now Turing is to be forgiven for having chosen symbol input/output capability as the paradigm for his T2, for at least three reasons: (1) computation is very powerful and general; it can simulate and approximate just about any other physical process. (2) Language, too, is very powerful and general; it can describe in words just about any object, action, event, state and trait. (3) Restricting T2 to email interactions rules out the irrelevant factor of physical appearance, which might bias our judgment: Turing's criterion is "cognition is as cognition does": cognitive capacity indistinguishable from our own. Not: "it must look just like any of us."

But some of our cognitive capacities do depend on things that cannot be tested by the standard "email" version of the Turing Test, T2 . People are able not only to name, describe and reason about objects, actions, events, states and traits in the world; they can also recognize, identify, manipulate and otherwise interact with them in the world. So the candidate has to be a robot, not just an email pen-pal. In fact, without the sensorimotor capacities of a robot, it is not clear how even the email T2 could be passed successfully: Would it not arouse immediate suspicion if our pen-pal was always mute about photos we sent via snail-mail? And is there any way to give a credible description of what anything feels like without ever having seen, touched, heard or felt anything?

So chances are that even to be able to pass the email version of the Turing Test, T2, the candidate would probably have to have, and draw upon, the capacity to pass the robotic version of the Turing Test too: Let's call robotic Turing Test T3. T2 – which is just language-in/language-out – simply draws on and tests T3 capacity indirectly. Perhaps Turing never meant that the candidate could only be a computer, computing. Computing is just the manipulation of meaningless symbols, based on rules operating on the shapes of the symbols. The shapes of the symbols are arbitrary in relation to the things that the symbols are interpretable (by those who understand what they mean) as referring to. Even for those who understand English, the string of symbols "the apple is red" does not *resemble* apples, red, or apples being red. For Searle, in the Chinese room, the same sentence, in Chinese, would not even *mean* "the apple is red." It would (as Searle insisted, memorably) only be a meaningless series of "squiggles" and "squoggles."

Symbol Grounding. What makes English meaningful to Searle? We know Searle can pass T2 in English; he can pass T3 in English too. And, for the reasons we've mentioned, it's unlikely that he could pass T2 in Chinese without also having the

ability to pass T3 in Chinese; and he cannot do that, because he has no idea what the symbols mean. In other words, the thought experiment of being able to pass T2 through symbol manipulation alone was probably a fiction all along: it can't be done.

But we still haven't said what makes English meaningful to Searle. His English symbols refer to objects, actions, events, states and traits in the world; and Searle can recognize, identify, manipulate and otherwise interact with those objects, etc., including being able to name, describe and reason about them – in English. Let's call that *know-how*: Searle – or, rather, Searle's brain -- has the know-how to pass T2 and T3 in English. Sensorimotor capacities are not computational but dynamic, so some of the underlying mechanisms producing this know-how must be dynamic (i.e., analog), rather than just computational (i.e., digital, symbolic). And hence cognition is not just computation.

Let us say that unlike his Chinese symbols, Searle's English symbols are "grounded" in his sensorimotor capacity to interact with the things in the world that his symbols refer to: they connect his words to their referents. Is sensorimotor grounding the same as meaning, then? Can we be certain that a T3-passing robot would really mean what it said?

The Cogito. Well perhaps certainty is a bit too much to ask. We already know from Descartes that we can be certain about the necessary, provable truths of mathematics but apart from that (with one prominent exception we will get to in a minute), there's no certainty. We can't even be certain about the laws of physics: They are just highly probably true. We can't be certain about the existence of the outside world; it's just highly probable. Same for the existence of other people, and for the fact that other people think: Highly probable, not certain. So what does that matter? Maybe certainty is something that one can only have in the formal world of mathematics. After all, things are *true* in the physical world of objects and people too, and we can know them; it's just that we can't know them *for sure*.

But Descartes also pointed out another certainty, one that would seem to be at the very opposite pole from the certainties of the abstract world of mathematics: We can also be sure that we think (Descartes' celebrated "Cogito"). It's impossible to doubt you're thinking, because doubting is thinking. That sounds like a trick, so let's put it in a more intuitive way: When I'm feeling something, I can't doubt that I'm feeling something. I'm feeling whatever I'm feeling. When I have a toothache, I can doubt that it's really my tooth that's ailing. Maybe it's referred pain from an eye infection. Maybe I have no tooth at all; it's been extracted. Worse, maybe it's really true that there's no outside world, and that my body and everything else is just an illusion, a dream! But what I can't doubt is that it *feels like* something when I'm feeling something: Whether or not I have an aching tooth, it feels like something (as it happens, a toothache), when I have a toothache. And that's a certainty. Feeling is a certainty (when you're feeling). Whenever you're feeling something, something is being felt, without a doubt. Whatever it feels like, that's what is being felt, without a doubt. Things may not *be* what they feel like (an injured tooth), but they certainly *feel like* whatever they like (a toothache), whilst they're being felt.

Meaning. Now back to the question: Is *grounding*, then, the same thing as *meaning*?

Can we be certain that a grounded T3-passing robot would really mean what it said (or that it would really be meaning anything at all)?

At the very least, we now know that if the robot does mean what it says, we can't be certain that it means what it says, for roughly the same reason that we can't even be sure there's a robot there, or an outside world. But let's set that aside as idle sceptical fretting.

Is it any worse, with the robot's meaning, than it is with the outside world existing, or the truth of the laws of physics? Is it just a matter of settling for high probability rather than insisting on certainty?

Turing seems to suggest that it is just that: The reason we should trust the TT is because it's no better or worse than what we have with one another: We can't be sure anyone else means what they say, or that they mean anything at all (or even that they exist). We can only be sure about ourselves – and even there, all I can really be sure of is that when I say something meaningful, it *feels* as if I know what I mean. I may be jabbering nonsense that is not at all what I mean to be saying.

Back to Searle in the Chinese room. He says he doesn't have any idea what he is saying when he manipulates the Chinese symbols. He does not know what (if anything) all of those symbols he is manipulating mean. Can we take him at his word? After all, to the Chinese speaker with whom he has been corresponding for 40 years in Chinese, he certainly looks and acts like someone who means to say what he says, and knows what he means.

But we've decided to give Searle the benefit of the doubt, for Cartesian reasons: Maybe Searle sometimes talks nonsense when he feels he's saying something meaningful. We can all understand that. We all do that sometimes. But what about the reverse: Can he be saying something meaningful when he feels he's talking nonsense? For 40 years straight, non-stop? Surely – and I really mean *surely* here – something is amiss if something like that happens.

If someone speaks in tongues – tongues that he says, honestly, he doesn't understand – and what he says in those tongues nevertheless makes consistent sense, then what we conclude, quite naturally, is that he must be suffering from multiple personality disorder: But multiple personality is a pathology, and the personalities usually only emerge one at a time. Searle, in contrast, has an explanation: "I'm just manipulating symbols according to rules; I have no idea what it means." And, moreover, his explanation is all true.

So we cannot escape the conclusion that Searle is the only sure arbiter of whether or not he understands or means anything, when he is transmitting and receiving Chinese. He can say, with Cartesian certainty, that he is not understanding or meaning anything at all.

For over three decades now, Searle's Chinese room argument has been debated in connection with the question of whether or not cognition is just computation. But here I want to refocus on the much harder question of whether or not meaning is just know-how. We are no longer talking about whether computation alone could pass T2, nor whether, if it could, this would mean that the candidate understood

and meant what it said. We are at T3 level, and the candidate is no longer just a computer, computing, but a robot, a dynamical system, doing not only computation, but also sensorimotor transduction, and perhaps a lot of other essential internal dynamic processes in between as well.

Unlike Searle, who tells us, honestly, that he has no idea what the squiggles he is receiving and sending refer to, the T3-passing robot shows us that he does know, by pointing out their referents, and interacting with the real world of objects, etc., indistinguishably from any of us. His words square with his deeds, just as any of ours do. So is Turing right that to ask for anything more than T3 is not only impossible, but unreasonable, since we have nothing more to go on when we are mind-reading one another either?

The Brain. But is there nothing more? Might there be (1) something more to what it is that we are testing for with the T3? And might there be (2) something more to test with than T3?

The answer to both (1) and (2) is yes. In testing know-how with T3, we are not just testing for know-how, we are testing for *meaning* (1). We realized that T2 (i.e., meaningless symbols in, meaningless symbols out, symbol-manipulation in between) was not even enough to test whether and how the symbols were grounded in their referents. The connection had to be made through the mediation of the mind of an external interpreter, whereas our robot was supposed to be making its own connections to its referents, autonomously. T3 fixed that. It grounded symbols in the robot's capacity for sensorimotor interactions with the referents of its symbols, at full T3 scale, for a lifetime. But does robotic indistinguishability from any of us mean total indistinguishability?

This is the point at which those who have been burning to bring the brain into this discussion all along can remind us that, after all, although we don't actually test it when we are mind-reading one another every day, the presumption is that what makes us all pretty much the same is not just that we behave indistinguishably from one another, but that we all have roughly the same brains (2). So there is, in principle, a Turing Test that is even more exacting than T3, and that is the neurobehavioral Turing Test, T4: The candidate must be totally indistinguishable from us not only in its verbal performance capacity (T2) and its sensorimotor performance capacity (T3) but also in its neurobehavioral performance capacity. After all, what the brain does internally is just as observable (especially today, in the era of brain imagery) as what the body does externally. Why would any good empirical scientist want to ignore observable data?

We will return to T4 in a moment. But first, regardless of which TT we use, having the capacity to pass the TT surely isn't an operational definition of having a mind. We are trying to infer something from having the know-how to pass the TT. What are we trying to infer? The TT itself is direct evidence of having that know-how, but that's all. What else is there, besides the know-how?

Feeling. We're back to Searle (and Descartes), and what only Searle can know, and that is that it *feels like* something to say something and mean something by it. Of course, it also feels like something – for a real person, like Searle - to say something meaningless; and indeed, it feels like something to do most of the things that we do

while we're awake and acting voluntarily. Any feeling will do, but here we use for our example the case of what it feels like to say something meaningful. Only the speaker himself can know for sure. And if he tells me that he means what he says, and that he understands what he means, I take him at his word (and I'm right, and it's true, of course), exactly as I do when he says he has a headache.

So the answer to the question "What else is there, besides know-how?" is that it's exactly the same sort of extra thing that there is, besides know-how, in the case of a headache. In T2, the know-how underlying a headache is simply to be able to state that you have a headache, and to make the rest of your discourse consistent with that fact. In T3, you might also have a pained expression on your face, cradle your head in your hands, and react in an uncharacteristically abrupt way when touched or spoken to. That's headache know-how too. If I suspect you are faking it, I could move to T4 and request a brain scan. (Let's pretend brain imagery is so advanced that it can reliably detect the neural correlates of a headache.) If the brain scan is positive, can I be sure you have a headache? As sure as I can be of other empirical truths, like apples falling down rather than up, F = ma, and there's a real world out there.

But is Turing right that if it's not Searle's headache that's in question, but the headache of a T3 robot, then I can be equally confident? Certainly not via T2. Is T3 enough? Or do I need T4? We normally only invoke brain scans and lie detectors with real people when there are reasons to believe they may be lying. The robot is T3-indistinguishable from us. Do I really need T4 to confirm he has a headache when he says he does, even if he otherwise behaves exactly like a person whom I had no reason to suspect was lying, and hence I wouldn't dream of ordering a brain scan every time he said he had a headache?

Never mind headaches. What about meaning? It is hard to imagine a brain scan for meaning, but suppose there was one. Suppose, for example, that when Searle said he could not understand the Chinese symbols that he was receiving and sending, despite the fact that he was exhibiting T2 know-how, a scanner could confirm that he was indeed exhibiting all the brain correlates of *not* understanding, rather than understanding.

But now suppose a robot – not Searle – passed T3 and failed T4. Better still, to make this even more realistic, suppose John Searle himself, the very one, in California, was discovered, when participating as a voluntary subject in his first cognitive neuroscience experiment, to be a T3 robot, and to have been one throughout his lifetime, as professor, relative, colleague and friend. How confident would his now failing T4 make us that he therefore had not meant anything he had said all his life? Would we be as confident as we were when the ostensibly human Searle had assured us, in the Chinese room, that he was *not* meaning anything when he communicated in Chinese? But in that hypothetical case, there was a plausible explanation for why Searle could not understand the Chinese. (He had never learned it, and was just manipulating squiggles and squoggles.) Is failing T4 like that?

Mind-Reading. This example is of course playing with our mind-reading intuitions. We could test them still further. Would failing T4 make us confident that Searle

could then immediately be dismembered, having turned out to be just a mindless device, perhaps to have his components studied by scientists, to reverse-engineer how they work, or to trace who had built him? How would Searle's family and friends feel about that?

Perhaps it's not fair to force us to make an intuitive or moral judgment based on such hypothetical examples, because there are no TT-scale robots – T2, T3 or T4 – and perhaps they're not even possible. Perhaps only a biological organism more or less like us in every respect could pass any of the Turing Tests, and if it could pass one, it could pass them all.

If that were true, then Turing's principle – cognition is as cognition does – would still be correct, but his research methodology would not be. The way to reverse-engineer human cognitive capacity would not be to try to build devices that can do what we can do, but to study the brain in the same way we studied the heart, kidneys and lungs – all biological organs with certain functions, by direct observation and manipulation.

The trouble is that the functions of the brain are *our* functions. What hearts, kidneys and lungs can do is mainly mechanical and chemical – pump blood or air, filter fluids, and so on – whereas what brains can do is what we can do. Their know-how is our know-how. And so far, computation and robotics have been the only ways we have derived even the vaguest inklings of how anything at all could do what our brains can do. Computational and robotic devices are so far toys, compared to us; but they are able to do a tiny fragment of what we are able to do. In contrast, neuroscience has not yet produced a causal explanation of anything that we can do (apart from "vegetative" functions such a temperature regulation, balance or breathing, which are more like the functions of the heart or kidney than the brain).

So even if it does turn out that the brain's way is the only way to produce our know-how, it is hard to see how we will be able to know how the brain succeeds in doing it if we don't build models that work the way we think the brain works, to test whether they can indeed do what the brain can do. And that begins to look more and more like the Turing Test again. Because the goal of cognitive science is, after all to give a causal explanation. And causal explanations have to be testable.

Causality and the Explanatory Gap. Let me close with some reflections on causality, by returning, as promised, to the notion of "mind over matter." Let's all confess that regardless of the formal position we may take on the "mind/body" problem, it *feels like* something to have a mind, to think, to cognize; and that something does not feel passive. Not only do my sensory experiences feel like something, but so do my motor experiences. And when it comes to voluntary actions, it feels like I do what I do because I *feel like* it. It feels like I'm somehow causing my actions by "willing" them. (Don't ask me what's causing my willing; I'm tempted to say "me" but not even Descartes knows what that really means; only what it feels like.)

Now this is not really a digression on the issue of free will. It is just a closing reflection on the causal role of feeling in cognition, and in attempts to explain cognition causally. It is undeniable that we feel (that's Descartes' Cogito again). This

essay has suggested that our inescapable uncertainty about whether T2, T3 or T4 successfully capture and explain cognition turns out to reduce to the question of whether causal explanations only capture and explain know-how, or they also capture and explain feeling.

I want to suggest that causal explan can only ever capture and explain know-how, and that (just as Turing suggested) there's no point in asking for or expecting more. Regardless of whether we arrive at our explanation via T2, T3, T4, or via the direct observation, manipulation and modeling of brain function, we will always be faced with the uncertainty of whether we have explained all of cognition, or just our know-how.

And I think I can pinpoint the reason why we cannot hope to do any better than that: Causal explanation accounts for how and why things happen as they do, and account for it causally. Causal explanation of cognition – whether it is based on designing a mechanism that turns out to successfully pass T2, T3, or T4, or on modeling what gives the brain its T4 capacity – will always be open to the usual sort of skepticism that we had agreed to ignore as not worth fretting about. And perhaps it is indeed not worth fretting about the fact that, at the end of the day, the successful total explanation of our know-how will always be equally compatible with the presence or the absence of feeling. For unless we are prepared to be telekinetic dualists, according a separate, unique causal power to feeling itself ("mind over matter") -- for which there is no evidence, only overwhelming evidence against it – there is no causal room in any model for feeling.

Yet, although it may be an illusion that some of the things I do, I do because I feel like it, it is certainly not an illusion that it *feels like* some of the things I do, I do because I feel like it. And that feeling is as real as the feeling that I have a toothache even when I don't have a tooth.

So whereas it may well be that our T2, T3 or T4 candidate really feels -- and surely real people with brains do -- nothing in the causal explanation of the T2, T3 or T4 know-how will explain how or why we feel. I don't think that this is because feeling is mystical, or even because the right causal explanation would not "capture" feeling. It's just that whereas causal explanation explains how it captures know-how, it cannot explain how it captures feelings. And whereas it is transparent why having T2, T3, and perhaps T4 know-how would be useful for the Darwinian survival machines that we all are, it is not at all apparent how or why having feelings would be.

Harnad, Stevan (1989) Minds, Machines and Searle. *Journal of Theoretical and Experimental Artificial Intelligence* 1: 5-25.

_____ (1990) The Symbol Grounding Problem *Physica D* 42: 335-346.

_____ (1992) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. *SIGART Bulletin* 3(4); 9-10.

_____ (1994) Levels of Functional Equivalence in Reverse Bioengineering: The

Darwinian Turing Test for Artificial Life. Artificial Life 1(3): 293-301.

_____ (1996) Computation Is Just Interpretable Symbol Manipulation: Cognition Isn't. Special Issue on "What Is Computation" *Minds and Machines* 4: 379-390

_____ (1995) Why and How We Are Not Zombies. *Journal of Consciousness Studies* 1:164-167.

_____ (2000) Correlation vs. Causality: How/Why the Mind/Body Problem Is Hard. [Invited Commentary of Humphrey, N. "How to Solve the Mind-Body Problem"] *Journal of Consciousness Studies* 7(4): 54-61.

_____ (2000) Minds, Machines, and Turing: The Indistinguishability of Indistinguishables. (special issue on "Alan Turing and Artificial Intelligence") *Journal of Logic, Language, and Information* 9(4): 425-445.

_____ (2001) Spielberg's AI: Another Cuddly No-Brainer.

_____ (2002) Turing Indistinguishability and the Blind Watchmaker. In: J. Fetzer (ed.) *Evolving Consciousness.* Amsterdam: John Benjamins. Pp. 3-18.

_____ (2005) What Is Consciousness? *New York Review* 52 (11)

_____ (2006) Searle's Chinese Room Argument. In: Borchert, Donald (ed) *Encyclopedia of Philosophy*. Macmillan Reference. Volume 2: 239-242

_____ (2008) What It Feels Like To Hear Voices: Fond Memories of Julian Jaynes. In: *Biennial Julian Jaynes Conference on Consciousness*, 7-9 August 2008, University of Prince Edward Island.

_____ and Scherzer, P. (2008) First, Scale Up to the Robotic Turing Test, Then Worry About Feeling. Artificial Intelligence in Medicine 44(2): 83-89

_____ (2008) The Annotation Game: On Turing (1950) on Computing, Machinery and Intelligence. In: Epstein, Robert & Peters, Grace (Eds.) *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer

_____ (2009) Cohabitation: Computation at 70, Cognition at 20, in Dedrick, D., Eds. *Cognition, Computation, and Pylyshyn*. MIT Press

_____ (2011) Zen and the Art of Explaining the Mind. *International Journal of Machine Consciousness* (IJMC)