First, Scale Up to the Robotic Turing Test, Then Worry About Feeling

Stevan Harnad[1,2] **& Peter Scherzer**[1]

[1]Cognitive Sciences Institute
Université du Québec à Montréal
Montréal, Québec, Canada H3C 3P8
[2]Electronics and Computer Science,
University or Southampton, UK SO17 1BJ
harnad AT ecs dot soton dot ac dot uk

Abstract

Consciousness is feeling, and the problem of consciousness is the problem of explaining how and why some of the functions underlying some of our performance capacities are felt rather than just "functed." But unless we are prepared to assign to feeling a telekinetic power (which all evidence contradicts), feeling cannot be assigned any causal power at all. We cannot explain how or why we feel. Hence the empirical target of cognitive science can only be to scale up to the robotic Turing Test, which is to explain all of our performance capacity, but without explaining consciousness or incorporating it in any way in our functional explanation.

Consciousness is Feeling. First we have to agree on what we mean by consciousness. Let us not mince words. To be conscious of something means to be aware of something, which in turn means to *feel* something. Hence *consciousness is feeling*, no more, no less. An entity that feels is conscious; an entity that does not feel is not. The rest is merely about *what* the entity feels. What it is feeling is what it is conscious *of*. And what it is not feeling, it is not conscious of (Nagel 1974; Harnad 2003).

A counterintuition immediately suggests itself: "Surely I am conscious of things I don't feel!" This is easily resolved once one tries to think of an actual counterexample of something one is conscious of but does not feel (and fails every time). The positive examples of feeling are easy, and go far beyond just emotions and sensations: I feel pain, hunger and fear. I also feel what it is like to see blue, hear music, touch wood, or move my arm. More subtle, but no less feelingful (hence no less conscious), is feeling what it is like to think (or to understand or believe or know) that that is a cat, that the cat is on the mat, that 2+2 = 4. To think something, too, is to feel something.

[Pause: Doesn't (1) knowing that that's a cat feel different from (2) not knowing whether that's a cat? or from (3) knowing that that's not a cat? If you think they all *feel* the same, then what *is* the difference between them – apart from the facts, about the cat? For the facts about the cat certainly don't determine whether I know them or not. And forget about "unconscious knowing": If I believe it's a rat, yet I treat it exactly as if it were a cat, then I believe it's a rat, but I treat it as it were a cat. And it feels like something to believe it's a rat, even if I treat it like a cat, and even it's really a cat! It feels like something else to believe it's a cat. So if I don't feel it's a cat, I don't believe it's a cat, let alone know it's a cat, no matter how I act. Freud slipped on this one, with his incoherent notion of an unconscious mind -- a zombie alter-ego. Unconscious knowing makes no more sense than unfelt feeling. Indeed it's the same thing. And unconscious know-how is merely performance capacity, not unconscious "know-that." But we are getting ahead of ourselves.]

Clearly to think and to know is also to have -- and usually also to perform in such a way as to be able to confirm having -- certain data and know-how. In order to know that that is a cat, I have to have to be able to identify it as a cat. But it is trivially easy to get a machine to identify something as a cat without feeling a thing. In that case, the identification is not conscious. So just having the data and the ability to act on it is not enough.

Performance capacity. Not enough for consciousness, but enough for performance capacity -- and not only is performance capacity what robotics is about, but it is also what cognitive science is about: Both robotics and cognitive science try to explain the causal basis for performance capacity, the functional mechanism that generates it: robotics, in order to get machines to do useful things for us, and cognitive science, in order to explain how we ourselves are able to do such things.

And *know-how* -- sensorimotor skill -- can come in conscious and nonconscious (i.e., felt and unfelt) form: When we do something (consciously), it *feels like something* to do it; if it did not feel like that, we would feel shocked. I would be dismayed to see my fist clench if I did not also feel that I was clenching it, and clenching it because I felt like clenching it. I could understand if my fist clenched because my doctor had hit a reflex point, or because I had a muscle disease (Cohen et al. 1994). But that would still feel like something: like my fist clenching, involuntarily (i.e., because of the reflex or the disease, not because I had willed it). Even if the involuntary spasm occurs while my hand is anaesthetized, I can see the spasm, and that feels like something. And even if my hand is behind a screen and anaesthetized, and someone simply *tells* me that my fist just clenched, *that* feels like

something (though that something is really only what it feels like to hear and understand and believe that my fist has clenched without my willing or feeling the clenching). If the clenching occurs while I'm in dreamless sleep or a coma, then it is unconscious, just as it would be in any of today's robots. The only difference would be that I can eventually wake up, or recover from the coma, and feel again, and even feel what it's like to hear and believe that I had been in a coma and that my fist had been clenched while I was in the coma (so I am told, or so the video shows me). No such possibility for today's robots. They don't feel a thing (Harnad 1995).

Biochemical Robots. By "today's robots," we mean the real robots we build today (not our fictional and cinematic ones, which are way ahead of the game: see Appendix). We are ourselves natural biochemical robots that were built by the Blind Watchmaker (i.e., Darwinian evolution; Harnad 2002). So what distinguishes us from today's man-made robots is not that we are not robots -- a "robot" is simply an autonomous sensorimotor system with certain performance capacities -- but that we happen to be robots with performance capacities that vastly exceed those of any robot built by us so far. And a further capacity – but not a performance capacity – that our current robots lack is the capacity to feel.

The real question, then, for cognitive robotics (i.e., for that branch of robotics that is concerned with explaining how animals and humans can do what they can do, rather than just with creating devices that can do things we'd like to have done for us) is whether feeling is a property that we can and should try to build into our robots. Let us quickly give our answer: *We can't, and hence we shouldn't even bother to try.*

The Other-Minds Problem. Justifying this answer takes a bit longer. First, it's not that there is any doubt at all about the reality of feeling in people and animals. Although, because of the "other-minds" problem, it is impossible to know for sure that anyone else but myself feels, that uncertainty shrinks to almost zero when it comes to real people, who look and act exactly as I do. And although the uncertainty grows somewhat with animals as they become more and more unlike me (and especially with one-celled creatures and plants), it is very likely that all vertebrates, and probably invertebrates too, feel (Harnad 1991).

The Causal Role of Feeling. So the problem is not with uncertainty about the reality of feeling: the problem is with the *causal role* of feeling in generating (and hence in explaining) performance, and performance capacity. Let us agree that *to explain something is to provide a causal mechanism for it.* The concept of force plays an essential explanatory role in current physical theory. Until/unless they are unified, there are four forces: electromagnetism, gravitation, and the strong and weak subatomic forces. There is no evidence of any further forces. Hence even when it feels as if I've just clenched my fist voluntarily (i.e., because I felt like it, because I willed it), the real cause of the clenching of my fist voluntarily has to be a lot more like what it is when my fist clenches involuntarily, because of a reflex or a muscle spasm. For feeling is not a fifth causal force. It must be piggy-backing on the other four, somehow. It is just that in the voluntary case it *feels as if* the cause is me.

But the other four forces are all unfelt forces. And the dynamical systems whose properties those forces are used to explain, causally (whether they are subatomic interactions, billiard ball collisions, clocks ticking, cars driving, plants growing, animals behaving, solar systems revolving or the Big Bang exploding) are all unfeeling systems -- with the exception of some animals (though probably not plants). Animals feel, but the question is: how and why do they feel? And the problem is to answer this question using only the known four forces, all of them unfelt forces.

The Mind/Matter Problem. The problem is the flip-side of the other-minds problem, and it is called the "mind/matter" (or "mind/body") problem. It had a precursor: the "life/matter" problem. We once thought it was impossible to explain life without a fifth "vital" force. But that turns out to have been wrong. Genetics, biochemistry, anatomy, physiology, and developmental and evolutionary biology are managing to explain all known properties of life using only the four known forces. But will those suffice to explain feeling? They no doubt suffice to *generate* feeling, somehow, but not to explain *how or why* they generate it -- and that is the mind/matter problem (Harnad 2000).

Forward and Reverse Engineering. In a sense, all of biology is reverse engineering: In forward engineering, we build artificial systems that do useful things (as in ordinary robotics) and in reverse-engineering we try to give a causal explanation of how an already-built system works (as in cognitive robotics). All biological systems were "built" by the Blind Watchmaker (evolution). So the explanatory task of biology is to reverse-engineer what evolution built, in order to explain how it works: functionally, causally. Often this requires building real or virtual models to test whether or not our causal explanations actually work (Harnad 1994).

Vitalism. In the case of the reverse-engineering of life itself, it turned out that no extra "vital" force was necessary to explain all the structural and functional properties of living matter. It is no longer even apparent today why anyone would ever have imagined that there might need to be a special life force, for there was never really any "life/matter" problem. The structure, function and I/O (Input/Output) performance capacities of biological systems are all perfectly objective, observable, and explicable properties, like all other physical properties. In contrast, with the "other-minds" problem, we each know perfectly well what it is that would be *missing* if others did not feel at all, as we do: feeling. But "living" has no counterpart for this: Other systems are alive because they have the objective, observable properties of living systems. There is no further unobservable property of "living" about which there is some additional uncertainty -- no property whose presence you can only ascertain by *being* the system, as in the case of feeling. (In fact, although they

may not have realized it, the vitalists were probably thinking of the mind/matter problem itself when they imagined that life was special, that it needed some sort of special life force. They were implicitly assuming that *living* matter had to be *feeling* matter.)

Insofar as cognitive robotics is concerned, what we have is *performing matter* that also happens to feel -- indeed feels as if it performs *because* it feels. I/O performance capacity itself is something objective and observable, hence functionally explicable. If we were all just feelingless Darwinian survival machines (as Darwinian biology would have predicted), the methodology and goal of cognitive robotics would be clear and unproblematic: reverse-engineering our performance capacities. This is essentially the Turing Test (Turing 1950; Harnad 1992), taken as both cognitive science's means and its end (Harnad 1995, 2007). What Turing's method appears to miss, however, is feelings. So it is only natural to ask whether there is any way to reverse-engineer feelings too, along with performance capacities.

Correlation and Causation. First, let us be sure to separate feelings from their functional correlates (Harnad 2000): We feel pain when we have been hurt and we need to do something about it: for example, removing the injured limb from the source of the injury, keeping our weight off the injured limb, learning to avoid the circumstances that caused the injury. These are all just adaptive nociceptive functions. Everything just described can be accomplished, functionally, by merely detecting and responding to the injury-causing conditions, learning to avoid them, etc. All those functions can be accomplished without feeling a thing; indeed, robots can already do such things today, to a limited degree. So when we try to go on to explain the causal role of the fact that nociceptive performance capacity's underlying function is a *felt* function, we cannot use nociception's obvious functional benefits to explain (let alone give a causal role to) the fact that nociceptive function also happens to be felt: The question persists: how and why?

The same is true of thinking and understanding: It is clear why it would be adaptive for a Darwinian survival machine to learn and plan -- and adaptive also for a social population of survival machines to have language, to speak, and to exchange useful information (Cangelosi & Harnad 2001). What is not clear is why any of that function should be *felt*, rather than merely "functed."

Can we not just satisfy ourselves, then, with feeling as a "correlate" of function? Can we not, by the very same commonsense means we use to settle the other-minds problem ("Surely if other human beings look and act just the same way I do, then they too are feeling, as I do, even though I cannot be absolutely certain that they are!") also settle the functional-correlates problem? "Surely the neural activity that accompanies pain *is* the pain, in some sense!"

Feeling Versus "Functing": How and Why Do We Feel? In some sense. But that is precisely what makes the mind/matter problem such a hard (probably insoluble) problem: Because we cannot explain *how* feeling and its neural correlates are the same thing; and even less can we explain *why* adaptive functions are accompanied by feelings at all. Indeed, it is the "why" that is the real problem. The existence of feelings is not in doubt. The "identity" of feelings with their invariant neural correlates is also beyond doubt (though it is also beyond explanatory reach, hence beyond comprehension). We are as ready to accept that the brain correlates of feeling *are* the feelings, as we are that other people feel. But what we cannot explain is *why*: Why are some adaptive functions felt? And what is the causal role -- the adaptive, functional advantage -- of the fact that those functions are felt rather than just functed?

Before we go on, let us note that this question would have profound implications for cognitive robotics if it in fact had an answer. If we could explain what the causal advantages of feeling over functing were in those cases where our functions are felt (the "why" question), and if we we could specify the actual causal role that feeling plays in such cases (the "how" question), then there would be scope for an attempt to incorporate that causal role in our robotic modeling. But if it turns out that we cannot make functional or causal sense of feeling at all, then cognitive robotics is just I/O performance capacity modeling (exactly as Turing said it was), and there is no point in trying to do anything with or about feeling.

Telekinesis. There are two reasons to be pessimistic about making feeling into a causal component in robotic modeling and cognitive explanation. One reason has already been mentioned: There is no evidence at all that feeling is or can be an independent causal force, even though it *feels as if* it is. For the clenching of my fist to be caused by my willing it to be clenched, rather than by some combination of the usual four feelingless forces of nature, would require evidence of a fifth causal force -- a telekinetic force. And there is no such evidence, hence no such fifth force.

The second reason comes from the neural correlates of voluntary action: If the neural correlates of felt intention were simultaneous with the functional triggering of voluntary movement in the brain, that would be bad enough (for, as noted, there would be no explanation at all for why intention was felt rather than just functed). But the situation may be even worse: The research of Libet (1985) and others on the "readiness potential," a brain process that precedes voluntary movement, suggests that that process begins *before* the subject feels the intention to move. So it is not only that the feeling of agency is just an inexplicable correlate rather than a cause of action, but it may come too late in time even to be a correlate of the cause, rather than just one of its aftereffects.

With all this in mind, we now turn to the six specific questions that were addressed to this AAAI Symposium on AI and Consciousness:

1. Are models of consciousness useful for AI? *No*. First, consciousness is feeling. Second, the only thing that can be "modeled" is I/O performance capacity, and to model that is to design a system that can generate that performance capacity. Feeling itself is not performance capacity. It is a correlate of performance capacity. The best

that AI can do is to try to scale up to full Turing-scale robotic performance capacity and to *hope* that the conscious correlates will be there too. If there is anything we learn about neural function, or the neural correlates of feeling, that can help AI generate the performance capacity, by all means use and apply it. But for now it is neuroscience that is looking to AI and robotics for functional mechanisms to help explain neural performance data and to help guide further neural data-gathering, not vice-versa.

2. Are AI systems useful for understanding consciousness? *Not at all.* They are useful only inasmuch as they help explain performance capacity. Everything pertaining to consciousness (feeling) will be merely a mentalistic interpretation of the functional mechanism of performance capacity (i.e., a hermeneutic exercise, rather than the causal, empirical explanation that is needed). And mentalistic interpretation will merely cover up the impoverished level of today's performance-capacity modeling.

3. What are the theoretical foundations of machine consciousness? *There are no theoretical foundations of machine consciousness.* Until further notice, neither AI nor neuroscience nor any other empirical discipline can even begin to explain how or why we feel. Nor is there any sign that they ever will.

4. Is machine phenomenology possible? *Only as an empty hermeneutic exercise* (merely overinterpreting our current generation of toy models by projecting a mentalistic interpretation onto them) – until we design a candidate that actually passes the Turing Test. Then there might be some realistic hope that it actually *has* a phenomenology (i.e., feelings). But we won't know whether it does: it alone will know whether it does. And even if it does feel, we won't be able to explain how or why it does.

5. Will conscious systems perform better than unconscious systems? *The question should have been the reverse*: Will systems that can perform more and better be more likely to feel? The answer to that might be a guarded yes, if we imagine systems that scale up from invertebrate, to vertebrate, to mammalian, to primate to human performance capacity, Turing-scale. We can be pretty confident that none of the systems we've designed so far even comes close to feeling. The system that passes the human Turing Test has the best chance, but even there, we won't know whether, how, or why.

6. What are the implementation issues of current AI systems inspired by consciousness? *There are no implementation issues inspired by consciousness.* There are just internal structures and processes that we overinterpret mentalistically. In neural models, measuring, predicting and manipulating the correlates of consciousness has some clinical value, but it still does not provide a hint of causal explanation. And consciousness in today's AI and robotic (and neural) models is purely decorative, not functional. At a time when performance capacity modeling is still so impoverished, mentalistic interpretations only cover up the yawning performance deficits. Think only of implementing what will generate more powerful performance capacity, and worry about consciousness only if and when you have generate our performance capacity, Turing scale.

Appendix

Spielberg's AI: Another Cuddly No-Brainer

It would have been possible to make an intelligent film about Artificial Intelligence -- even a cuddly-intelligent film. And without asking for too much from the viewer. It would just ask for a bit more thought from the maker.

Spielberg's (2001) AI is about a "robot" boy who is "programmed" to love his adoptive human mother but is discriminated against because he is just a robot. Both "robot" and "programmed" are in scare-quotes, because these are the two things that should have been given more thought before making the movie. [Most of this critique also applies to the short story by Aldiss (1969) that inspired the movie, but the buck stops with the film as made, and its maker.]

What Is/Isn't a Robot? So, what is a "robot," exactly? It's a man-made system that can move independently. So, is a human baby a robot? Let's say not, though it fits the definition so far! It's a robot only if it's not made in the "usual way" we make babies. So, is a test-tube fertilized baby, or a cloned one, a robot? No. Even one that grows entirely in an incubator? No, it's still growing from "naturally" man-made cells, or clones of them.

What about a baby with most of its organs replaced by synthetic organs? Is a baby with a silicon heart part-robot? Does it become more robot as we give it more synthetic organs? What if part of its brain is synthetic, transplanted because of an accident or disease? Does that make the baby part robot? And if all the parts were swapped, would that make it all robot?

I think we all agree intuitively, once we think about it, that this is all very arbitrary: The fact that part or all of someone is synthetic is not really what we mean by a robot. If someone you knew were gradually replaced, because of a progressive disease, by synthetic organs, but they otherwise stayed themselves, at no time would you say they had disappeared and been replaced by a robot -- unless, of course they did "disappear," and some other personality took their place.

But the trouble with that, as a "test" of whether or not something has become a robot, is that exactly the same thing can happen without any synthetic parts at all: "Natural" brain damage can radically change someone's personality, to the point where they are not familiar or recognizable at all as the person you knew -- yet we would not call such a new personality a robot. At worst, it's another person, in place of the one you once knew. So what makes it a "robot" instead of a person in the synthetic case? Or rather, what -- apart from being made of (some or all) synthetic parts -- is it to be a "robot"?

Programming. Now we come to the "programming." AI's robot-boy is billed as being "programmed" to love. Now exactly what does it mean to be "programmed" to love? I know what a computer program is. It is a code that, when it is run on a machine, makes the machine go into various states -- on/off, hot/cold, move/don't-move, etc. What about me? Does my heart beat because it is programmed (by my DNA) to beat, or for some other reason? What about my breathing? What about my loving? I don't mean choosing to love one person rather than another (if we can "choose" such things at all, we get into the problem of "free will," which is a bigger question than what we are considering here): I mean choosing to be able to love -- or to feel anything at all: Is our species not "programmed" for our capacity to feel by our DNA, as surely as we are programmed for our capacity to breathe or walk?

Let's not get into technical questions about whether or not the genetic code that dictates our shape, our growth, and our other capacities is a "program" in exactly the same sense as a computer program. Either way, it's obvious that a baby can no more "choose" to be able to feel than it can choose to be able to fly. So this is another non-difference between us and the robot-boy with the capacity to feel love.

Xenophobia. So what is the relevant way in which the robot-boy differs from us, if it isn't just that it has synthetic parts, and it isn't because its capacity for feeling is any more (or less) "programmed" than our own is?

The film depicts how, whatever the difference is, our attitude to it is rather like racism or xenophobia. We mistreat robots because they are different from us. We've done that sort of thing before, because of the color of people's skins; we're just as inclined to do it because of what's under their skins.

But what the film misses completely is that, if the robot-boy really can feel (and, since this is fiction, we are meant to accept the maker's premise that he can), then mistreating him is not just *like* racism, it *is* racism, as surely as it would be if we started to mistreat a biological boy because parts of him were replaced by synthetic parts. Racism (and, for that matter, speciesism, and terrestrialism) is simply our readiness to hurt or ignore the feelings of feeling creatures because we think that, owing to some difference between them and us, their feelings do not matter.

Now you might be inclined to say: This film doesn't sound like a no-brainer at all, if it makes us reflect on racism, and on mistreating creatures because they are different! But the trouble is that it does not really make us reflect on racism, or even on what robots and programming are. It simply plays upon the unexamined (and probably even incoherent) stereotypes we have about such things already.

There is a scene where still-living but mutilated robots, with their inner metal showing, are scavenging among the dismembered parts of dead robots (killed in a sadistic rodeo) to swap for defective parts of their own. But if it weren't for the metal, this could be real people looking for organ transplants. It's the superficial cue from the metal that keeps us in a state of fuzzy ambiguity about what they are. The fact that they are metal on the inside must mean they are different in some way: But what way (if we accept the film's premise that they really do feel)? It becomes trivial and banal if this is all just about cruelty to feeling people with metal organs.

"TT". There would have been ways to make "AI" less of a no-brainer. The ambiguity could have been about something much deeper than metal: It could have been about whether

other systems really do feel, or just act as if they feel, and how we could possibly know that, or tell the difference, and what difference that difference could really make -- but that film would have had to be called "TT" (for Turing Test) rather than "AI" or "ET," and it would have had to show (while keeping in touch with our "cuddly" feelings) how we are exactly in the same boat when we ask this question about one another as when we ask it about "robots."

Instead, we have the robot-boy re-enacting Pinnochio's quest to find the blue fairy to make him into a "real" boy. But we know what Pinnochio meant by "real": He just wanted to be made of flesh instead of wood. Is this just a re-make of Pinnochio then, in metal? The fact that the movie is made of so many old parts in any case (Wizard of Oz, Revenge of the Zombies, ET, Star Wars, Water-World, one couldn't possibly count them all) suggests that that's really all there was to it. Pity. An opportunity to build some real intelligence (and feeling) into a movie, missed.

References

Aldiss, B. (1969) Super-Toys Last All Summer Long. *Harper's Bazaar*, December 1969.
http://www.wired.com/wired/archive/5.01/ffsupertoys_pr.html

Cangelosi, A. and Harnad, S. (2001) The Adaptive Advantage of Symbolic Theft Over Sensorimotor Toil: Grounding Language in Perceptual Categories. *Evolution of Communication* **4**: 117-142.
http://eprints.ecs.soton.ac.uk/7709/

Cohen, H., S. Bouchard, P. Scherzer and H.A. Whitaker (1994) Language and Verbal Reasoning in Parkinson's Disease. Neuropsychiatry, *Neuropsychology and Behavioral Neurology* **7**:166-175.

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* **1**: 43-54.
http://eprints.ecs.soton.ac.uk/3379/

____ (1992) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. *SIGART Bulletin* **3**: 9 - 10.
http://cogprints.org/1584/

____ (1994) Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. *Artificial Life* **1**: 293-301.
http://eprints.ecs.soton.ac.uk/3363/

____ (1995) Why and How We Are Not Zombies. *Journal of Consciousness Studies* **1**:164-167.
http://cogprints.org/1601/

____ (2000) Correlation vs. Causality: How/Why the Mind/Body Problem Is Hard. *Journal of Consciousness Studies* **7**: 54-61.
http://cogprints.org/1617/

____ (2002) Turing Indistinguishability and the Blind Watchmaker. In: J. Fetzer (ed.) *Evolving Consciousness* Amsterdam: John Benjamins. pp. 3-18.
http://cogprints.org/1615/

____ (2003) Can a Machine Be Conscious? How? *Journal of Consciousness Studies* **10**: 69-75.
http://eprints.ecs.soton.ac.uk/7718/

____ (2007) The Annotation Game: On Turing (1950) on Computing, Machinery and Intelligence. In: Epstein, Robert & Peters, Grace (Eds.) *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Kluwer.
http://eprints.ecs.soton.ac.uk/7741/