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

Stevan Harnad & Peter Scherzer
Cognitive Sciences Institute
Université du Québec à Montréal
Montréal, Québec, Canada H3C 3P8
http://www.crsc.uqam.ca/en/index2_en.html

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 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.

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 (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 example (and fails). The 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, 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.

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. In order to know that that is a cat, I have to have the capacity to identify it as a cat. But it is trivially easy to get a machine to identify something as a cat without feeling anything. In that case, the identification is not conscious. So 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 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, 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. 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 intended 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 it 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 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.

Biochemical robots. By "today's robots" I mean the robots we build today. We are ourselves natural biochemical robots that were built by the Blind Watchmaker (evolution); 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 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 in 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.

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" 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.

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.

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 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. (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 a 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 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-engineer our performance capacities. This is essentially the Turing Test, taken as both cognitive science's means and its end (Harnad 2006). 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: 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 remains: 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. 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 potential 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.

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 the 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-Test scale robotic performance capacity and to hope that the 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.

Are AI systems useful for understanding consciousness? Not at all. They are useful only inasmuch as they help to explain performance capacity. Everything pertaining to

consciousness (feeling) will be merely interpretation (i.e., a hermeneutic exercise, rather than the causal, empirical explanation that is needed), and will merely cover up the impoverished level of performance-capacity modeling.

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 explain how or why we feel.

Is machine phenomenology possible? Only as an empty hermeneutic exercise (merely overinterpreting our current generation of toy models by projecting a mentalistic interpretation on 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, and, even if it does, we won't be able to explain how or why it does.

Will conscious systems perform better than unconscious systems? The question should have been the reverse: Will systems that 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.

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. Consciousness in today's AI and robotic models is purely decorative, not functional. At a time when performance 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 performance capacity licked, Turing scale.

REFERENCES

Harnad, S. (2003) Can a Machine Be Conscious? How? *Journal of Consciousness Studies* 10(4-5): 69-75. http://eprints.ecs.soton.ac.uk/7718/

Harnad, S. (2006) 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/

Libet, B. 1985. "Unconscious cerebral initiative and the role of conscious will in voluntary action". *Behavioral and Brain Sciences* 8: 529-566.

Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83: 435 - 451.