To appear in McCarthy, N. (ed.), *Philosophy of Engineering: Proceedings of a Series of Seminars held at The Royal Academy of Engineering*. London: Royal Academy of Engineering, 2009.

Interactive Empiricism: The Philosopher in the Machine

Ron Chrisley

COGS/Department of Informatics, University of Sussex

I would like to discuss some thoughts I've had recently about the general question we have been considering, that of the relationship between philosophy and engineering, and to present some possible new directions for collaboration between the two fields.

Take-home message

The conclusion I am heading for is this: I think there can be a two-way beneficial interaction between philosophy and engineering. Two qualifications can be made at the outset. First, I am sure there are possible collaborations other than the particular two-way interaction I'll be looking at. Second, a proper evaluation of the proposed interaction would include an investigation into the history of engineering to see if one can find examples of this two-way interaction, both to illustrate the interaction, and as a kind of validation of the fruitfulness of the interaction. I have not yet conducted such an investigation. But even if there have not yet been any interactions between philosophy and engineering of the form I have in mind, I maintain that the possibility is worth consideration, particularly because such interactions may be necessary for some philosophical breakthroughs and engineering achievements.

One of the more contentious claims I make to allow room for the mode of interaction I have in mind is this: some philosophical breakthroughs can only come about (or at least they are much more likely to occur) if philosophers, or the people who are struggling with the relevant conceptual questions engage in engineering; that is, design, build and interact with working systems appropriately related to the questions being considered. This is contentious in that most philosophers seek to draw a sharp line between *a priori* and *a posteriori* enquiry, with the discipline of philosophy entirely on the former side of the divide. On such a view, there is no room in philosophical methodology for scientific inquiry, even if it reliably yields knowledge of the world: the truths it discovers are empirical, whereas the truths

philosophy seeks are not, nor can they be established by consideration of such. However, the results of scientific enquiry (empirical truths) at least have the same format (propositional form) as the results of a priori inquiry, the universally accepted *modus operandi* of the true philosopher. All the more irrelevant, then, does engineering seem to philosophy. Not only does it traffic in the contingent and particular rather than in the necessary and universal, but it also fails to be a mode of enquiry at all, in the sense of a process that yields propositional knowledge.

I propose that this view is a misunderstanding of what philosophy is, or at least what it can be. It is only quite recently that such a sharp distinction has been made between philosophical and empirical forms of enquiry. I suspect many philosophers of the past, possibly including Kant, and definitely including Vico, would be sympathetic to the idea that engaging in engineering can lead to philosophical advances. Perhaps they had insights we would do well to recover.

There is another way that philosophy and engineering can interact, that is operates in a direction opposite to the interaction just mentioned; not by engineering helping philosophy, but by theorists/philosophers themselves being components that contribute to the dynamics of working systems. It may seem a bit odd, but what I will propose is that for the case of some complex systems, for instance an artificial consciousness, it might be necessary to incorporate a theorist or philosopher into the design. That might not make much sense right now, but later I will illustrate how this is possible by giving an example from actual work in artificial intelligence where I think this is a good way of understanding what is going on.

Thus, the incorporation of a philosopher or theorist into the design of a system has two aspects. One can consider the effect that the system dynamics will have on the theorist or philosopher. One can also consider the effect that the theorist/philosopher has on the artefact's dynamics; I will give a concrete example of that from some work at MIT.

Direction 1: Engineering conceptual change

The first direction, that of making philosophical progress by doing engineering, has to do with conceptual change.

One way of understanding philosophy is that it is about trying to solve conceptual problems. That this is not always appreciated by everyone is at the heart of

a joke I was told long ago by Brian Cantwell Smith. Some people, he said, make fun of philosophers for having struggled with very simple questions for millennia without having come up with an answer. For instance, consider the old chestnut: if a tree falls in a forest and no one is around to hear, does it make a noise? Some criticise philosophers for still pondering that question after all this time. But on the other hand if you give that question to scientists, they'll scratch their heads for a while, go away, write some things down on paper and then come back and say, 'Well, we've worked it out for elm and birch but we're still trying to solve the general case.' The relevance of the joke in this context is that the kind of answer the scientists gave is a sign that they didn't really understand the question; they have mistaken a conceptual problem for an empirical problem. Philosophy is about trying to resolve these conceptual problems. The orthodox way to solve such problems is through conceptual analysis, *a priori* enquiry, as discussed before. Nothing that I am going to say implies that such enquiry should not continue; I join the vast majority of philosophers in my conviction that not all limitations on our understanding, even scientific understanding, of the world are merely a matter of not having enough data. In particular, our problems in trying to understand consciousness are conceptual; the obstacles we face in understanding what it is for a physical thing to also be an experiencing thing aren't just a matter of not having enough knowledge of the nervous system.

Even if we knew much more about the nervous system than we do now, we would still have some fundamental puzzling questions. We have a naturalist intuition that consciousness, like anything else, is at root physical. Indeed, I am assuming here a broadly physicalist perspective: the belief that every kind of phenomenon in the world is grounded in physical happenings. On the other hand, we have another intuition – the philosopher Dan Dennett calls it the "zombic hunch" – that it is possible for there to be a creature – a zombie – that is physically just like us, but "there's no one home": it isn't conscious. At least some people believe that it is not inconceivable that there could be someone who is physically (and thus behaviourally) identical to you and yet different from you with respect to its experiential properties, even to the point of not having any experiences at all: a zombie-you. Those two intuitions are in direct conflict: the naturalist intuition is that like everything else, consciousness must be, at root, a physical phenomenon, whereas the zombic hunch implies that even if you fix in the physical you still haven't fixed the experiential.

The presence of both of these intuitions produces an unstaisfactory cognitive dissonance.

One way of responding to this is to diagnose the cognitive dissonance as the result of flaw in our concept of consciousness. If our concept of consciousness has paradoxical implications, perhaps we should try to develop a new concept of consciousness that doesn't. Perhaps we should look to conceptual change as a way to resolve this problem of the conflict between our naturalistic inclination and the zombic hunch.

***Conceptual* conceptual change?**

The suggestion that we solve our conceptual problems by changing our concepts is rather facile; it immediately prompts the question: how can we do this? One constraint is this: we don't want to change the subject. We want to change our concept of consciousness in a way that ensures that in employing the new concept, it is consciousness that we are still talking and thinking about; it is just that we are doing so in a better way. When we think about gold using our concept of gold, we think about the same thing that the ancients thought about when they thought about gold using their concept of gold. But we are not employing the same concept they did. We have a better concept of gold; we know what the essence of gold is – having an atomic number of 76. Although the ancients were confused and had many false beliefs about gold, it didn't mean they weren't thinking about gold. In fact, we can only make sense of their beliefs being false after we first understand them as being about gold.

Can we do the same thing for consciousness; can we refine our concept of consciousness? I propose that we can, and that we need to do so in order to solve some of the conceptual problems we face. But it seems unlikely that the kind of conceptual change required can itself come about through merely conceptual processes. By conceptual processes, I mean such processes as adding propositions to, or subtracting propositions from, one's stock of beliefs, whether it be by learning some more facts about consciousness or about the brain, or by engaging in philosophical arguments. Also, creating a new concept out of logical combinations of the concepts one already possesses. Such methods seem unable to surmount the impasse we have reached. If the recent history of discussions and disputes in the

areas of consciousness studies and the philosophy of consciousness is any indication, there is no rational way to convince somebody who has the zombic hunch to not have the zombic hunch, and vice versa (Dennett 2005). I am sceptical that the kind of conceptual change required can be achieved simply by reading journal articles about consciousness and other purely linguistic, propositional modes of inquiry. Don't get me wrong: of course these modes are essential to developing our understanding of anything, including consciousness. But it seems to me that there is reason to believe that they are not enough to resolve certain intractable conceptual problems, especially in the case of consciousness.

Non-conceptual conceptual change

If we are going to change our concept of consciousness so that we can make scientific sense of consciousness and understand the place of consciousness in the natural world, we might have to have our concept of consciousness undergo what you might call a non-conceptual change, a non-conceptual development of our concept of consciousness. What do I mean by non-conceptual development of a concept? I mean changes to a concept that aren't simply a case of adding or subtracting propositions to one's stock of beliefs, nor a case of creating a new concept out of logical combinations of the concepts one already possesses. But I am not concerned with just *any* kind of non-conceptual change to one's concept. Getting hit on the head, for instance, might change what you think or what you think you think about consciousness, or undergoing neurosurgery, or perhaps taking certain kinds of psychoactive drugs. Perhaps you will have an 'aha' experience if you do some of those things; perhaps some of them will result in non-conceptual conceptual change. But these are not, you may be happy to hear, the kinds of non-conceptual change I have in mind. The methods of change I do have in mind, like the bad examples just given, can't be achieved by hearing a philosophical argument, or by reading a passage of text. But unlike those bad examples, the changes are still *rational* changes that are *justified*, and in particular are *based on the experience of the subject matter*. Unlike the bad examples, where the change in your conceptual state is non-conceptual but random and not justified in any way, I am looking for ways that we can change our concept and yet have it be a change that tracks reality in some way, even if it can't be summarised in some text, or even if it can't be transmitted through text. This might sound impossible, but I think it is actually commonplace.

Concepts as skills

To see why this might be so, we need to get a clearer view of what concepts are. Consider the famous duck-rabbit figure, or the Necker cube. Wittgenstein argued that what underlies being able to move between the different ways of seeing the Necker cube or the duck-rabbit is the "mastery of a technique" (Wittgenstein 1972, p 208). This is exactly the kind of ability we are looking for in the case of conceptual change: the capacity both to see something in a new way and to see how it is the very same thing as what you saw in the old way. To be able to see how a thing that is appreciable from the experiential perspective is also the very same thing that is understandable from the physical perspective, to move seamlessly between those two viewpoints, is a skill. If so, then we now know that what we need in order to resolve the conceptual problems of consciousness is a skill: Acquiring the right concept of consciousness is a matter of acquiring a skill.

Note that skills are usually such that they can't transmitted through text alone. I can't give you a piece of text and thereby give you the ability to ride a bicycle. We can't have a philosophical argument that will give you that skill. Rather, skill in a domain typically requires experience of that domain. For instance, some symbolic, linguistic, propositional, information (the advice of a friend) helped me learn to juggle. But it wasn't sufficient for me to acquire that skill. It's true that I tried to acquire the skill without that conceptual knowledge and didn't do very well; it was only when my friend helped me, through language, to draw attention to certain aspects of my experience, I was able to juggle, in my own feeble way. But the advice alone didn't give me the skill; I had to attempt to juggle in order for the advice to be of any use. So also for the case of understanding consciousness. I am not saying there is no role for argumentation, thinking, reading journal articles, etc. I am only saying that these are probably not enough, and we need something else; we need a skill that cannot be acquired conceptually.

Interactive empiricism

The ideas that skill acquisition is a way to achieve non-conceptual conceptual change is an important part of a view that I call "interactive empiricism". This is distinct from empiricism *simpliciter*, that *all* concepts must be grounded in sense experience. Interactive empiricism is not a species of empiricism in that sense.

Rather, it is the view that the possession of *some* concepts requires having a particular kind of sense experience, a kind usually not emphasized in traditional empiricism. The "interactive" in "interactive empiricism" indicates what this particular kind of sense experience is: the sense experience involved in *interacting* with the subject matter of the concept, the stuff the concept is *of*. To have the kind of concept that will solve our conceptual problems, one must master a technique of understanding how one's perspective on the subject matter – in this case consciousness itself – will change in the light of one's different ways of intervening in that subject matter. Both the application and acquisition of this skill rerquire the having of experiences in the context of interaction with the subject matter of consciousness. Riding a bicycle isn't merely a passive reception of input of the kind the traditional empiricists were thinking of when they talked about grounding ideas in experience. The experience is interactive, the result of a dynamic engagement with the world. One acquires the skill of riding a bicycle because one experiences sensory feedback *in response to* one's actions; the kind of experience goes beyond a mere a one-way input from the world to your ideas.

To make that a little more clear, I want to draw attention to the fact that at least some movements in cognitive science are finding that interaction is essential to understanding cognition. Interaction is essential to perception on some views, such as O'Regan and Noë's sensory-motor contingency theory (O'Regan and Noë 2001). On that account you can only be perceiving the world if you have some capacity to interact with it, or if you are actually interacting with it. Consciousness in general is thought to involve interaction on views such as Susan Hurley's; hence the title of her book, *Consciousness in Action* (Hurley 1998). Cognition in general is thought to involve interaction essentially on some views such as Mark Bickhard's; hence the title of his "Interactivist Manifesto" (Bickhard 2008). A nice illustration of a concrete way in which interaction is crucial to certain kinds of cognitive development, in this case in visual perception, is the classic study by Held and Hein (Held and Hein 1963). They placed neonate kittens with undeveloped visual systems in an apparatus consisting of a circular room with a bar suspended from the ceiling, able to rotate about its midpoint. At the end of each bar is a harness for a kitten. The harnesses are such that one kitten is touching the ground and is able to move around relatively freely, but the other kitten is suspended in a way that its movements will not change

its position in the world at all; rather, its position is determined by the movements of the other kitten, which is able to walk around more or less normally. For the first kitten, there is a very natural interdependence between its actions and the visual input it receives. For the second kitten, there is little or no correlation between its muscle movements and the visual input it receives, because the visual input is largely determined by the first kitten. No matter what the second kitten does with its limbs, it doesn't, in the main, affect the input it receives. The result of this study is striking; after the developmental period; the first kitten has more or less normal vision, while the second kitten is more or less blind. This shows that having the right kind of interaction, engaging in action and having sense experience that is appropriately related to those actions, is crucial to certain kinds of development.

Meta cognitive science: Theorist as subject

Perhaps our conceptual development shares this property with visual development in kittens. If this is a general cognitive principle that governs our conceptual development as well, then we some developments in our concepts, for instance our concept of consciousness, may also require us to intervene in a subject matter and then receive reciprocal experiences that are appropriately related to those interventions. In the case we are considering, the interventions will in the phenomenon of consciousness itself.

A general science of human cognition should apply to individual cognizers; specifically, it should apply to cognitive scientists, philosophers, and engineers. And if one's cognitive science says that cognition in general, and conceptual development in particular, is interactive, it may also be that making philosophical advances via conceptual development will necessarily involve engaged, experiential activity.

Engineering as interaction

This is where engineering comes in. The kind of interaction that is relevant to understanding consciousness, cognitive systems, artificial consciousness, et cetera, is the design and construction of actual working systems that model or exhibit consciousness-related phenomena. It seems to me that these kinds of interactions will be the kind that will allow this conceptual development, that allow us to acquire the skills that constitute a conceptual advance with respect to consciousness.

I don't need to be so restrictive here; I don't need to deny that there might be other kinds of interaction that could assist in conceptual development with respect to consciousness. For instance, in interacting with each other or interacting with subjects in the experimental laboratory, if we were not only interacting in a more or less normal way, but also had access to real-time scans of each other's brains, this might also be a way of having the kind of interactive experience required to develop our concept of consciousness. But this is a much less plausible idea than the engineering based approach that I am suggesting here. First, there are the ethical issues: true interaction would require intervention on the lowest, neural level: directly altering another's neural state. But there are non-ethical problems as well. Compare trying to understand, say, how a computer works in a similar way, that is while you are interacting with the computer you have an oscilloscope and you can see what is going on in the lowest level hardware level of the computer while you are typing into Microsoft Word or something. In theory maybe you could get some great insights, and acquire some skill that would constitute a better concept of computation, but it seems unlikely; it is just too much of a jump from the lowest to the highest level to expect you to see some interesting correlations. There needs to be a step by step, level by level, structured approach that will allow interaction to have an effect on our concept of consciousness. So also, then, with the brain scan suggestion. By contrast, designing and building cognitive systems can and does have this mediated structuring of activity, and is thus more likely to be the kind of interaction that is going to yield the right kind of conceptual change.

(An aside for those familiar with the Knowledge Argument (Jackson 1982): Once one realises that science is itself an interactive activity that involves experiencing the world, the whole premise of Jackson's argument is revealed to be contradictory. Jackson asks us to imagine that Mary knows everything science has to say about colour vision but has never seen red. On the view of science presented here, this is revealed to be a contradiction. As Alter (Alter 1998) has independently observed, Jackson assumes that all the knowledge of physical science can be written down and acquired by Mary through reading. If what I have said here is right, then this assumption of Jackson's is false. Science in general involves the having of experiences, and colour science in particular involves the having of colour experiences. So for Mary to truly know everything that science knows about colour

vision, she would have to have all the experiences that scientists have needed to have about colour vision, in particular the experience of seeing red. To suppose this and to also suppose that Mary has not had the experience of seeing red is a contradiction.)

Direction 2: We are a part of the systems we build

To close, I would like to say a few things about the other direction of collaboration between philosophy and engineering. The key observation here is that we are part of the systems we build, and just as interaction can have a salutary effect on the philosopher, as just discussed, so also it could be that the philosopher or the theorist might be a crucial component in the developmental dynamics of that system. Not only is this an abstract possibility; it is also a concrete actuality in that some research in robotics is exploiting this means of interaction. To introduce this research, let me ask: What is the biggest engineering advance in artificial intelligence in the last twenty years? A provocative answer is: Kismet's eyebrows (Breazeal and Scassellati 2000). Putting eyebrows on the robot named Kismet may very well be one of the biggest technological and conceptual advances in artificial intelligence in the last twenty years. Let me explain. Kismet needed to learn how to track visual objects. It could only do that if the stimuli were appropriate; that is if they were moving within a certain range of speeds at a certain distance so Kismet could focus on them, etc. One could try to ensure that the trainer kept the stimuli within this range by a number of means. But a very efficient way of getting the trainer, a person, to keep the system within a particular part of the phase space was to put eyebrows on Kismet, and maybe adding a little more like having Kismet jump back and raise its eyebrows under certain conditions. When the state variable moved out the optimal region of phase space, Kismet jumped back and raised its eyebrows. Given our inbuilt dispositions to respond to such situations, one doesn't have to tell the trainer what to do in such a situation; one doesn't have to give them instructions, they don't have to consult a rulebook or anything. The trainer will just respond naturally, because we are built to respond to displays of "startlement" in particular ways, and the trainer will be non-conceptually disposed to treat raised eyebrows and pulling back as a display of startlement. The fact that Kismet is in fact not an experiencing creature, and therefore unable to be startled, is irrelevant. What is relevant is that Kismet behaves like an experiencing, startled creature, and therefore the trainer will respond unreflectively in the appropriate way. That is, the trainer will pull back, will move the stimulus back

into a proper part of the phase space with the result that tracking and learning will continue. That is a very efficient way to exploit the dynamical relationship between Kismet and the trainer in order to get Kismet to learn in the proper way. That is an example of a way in which the trainer, the theorist, the philosopher even, can be in the loop, be part of the system.

Combining the directions of interaction

These two directions of interaction can be combined. If we are part of the system, not only can the theorist/philosopher have a beneficial causal effect on the robot's performance, as we saw with Kismet; but, as we saw in the first part of the talk, interactions with the robot can also have a beneficial effect on the theorist/philopher by prompting conceptual change. This suggests an alternative design strategy for artificial consciousness. Instead of trying to design a machine consciousness in one step, we could instead see the design as a dynamical developmental process. On such a view, the first step is to design a system S1 in such a way that it will prompt relevant conceptual changes in us, such that those conceptual changes will allow us to design another system S2, so that S2 will prompt further conceptual changes in us, and so on. On this view, we see ourselves in a dynamical, a dialectical relationship with the systems we build and try to get on that trajectory, get onto that design spiral, rather than trying to get to the endpoint all in one go. If we think about this development, this design trajectory and apply some of the techniques of engineering to that trajectory, perhaps we will be able to get further in the quest for machine consciousness than we have been able to so far.

Frank Herbert was a prescient author; he wrote about this possibility in the 1960s in a novel entitled *Destination: Void* (Herbert 1966). In the novel, people attempt to create machine consciousness in an indirect way. First, they genetically engineer clones to have the right sort of skills and personalities to form a team which might make inventing machine consciousness more likely. These clones are then put into a carefully engineered technological environment, which included certain kinds of computing technologies and neural wetware, but which was located on a spaceship. Then these clones are manipulated and given certain kinds of motivations; specifically, they discover that the ship they are on is going to fail if they didn't create a machine consciousness first. The hope of the designers of the whole clone/spaceship/ hardware system is that if those ingredients are put together in the

right way, the clones will come up with a design for a machine consciousness, or at least come up with the next stage of such a design, which can be the starting point for the next generation of clones.

Thus, Herbert already anticipated the idea that including the engineer/theorist/philosopher into the design of the system might be essential for the construction of machine consciousness. Another interesting point is that a crucial part of the project in the novel is that the challenges the clone crew face are such that they are forced to think about what they mean by the word "consciousness". The engineered crises force the crew to engage in philosophy, and attempt to come up with a definition of consciousness. The crises are such that the crew can only see what consciousness is by being confronted with the embodied exigencies of the crises. The people designing this experiment don't know what consciousness is, but are rather hoping that they have assembled a conjunction of crew, situation and technology that will allow advance toward a solution to be made; "designing for emergence".

Although Herbert's work is mere science fiction novel, perhaps it isn't so far off from what we are or could be doing now. I don't just mean Kismet, although that is a good example, and I praised it as being a substantive breakthrough in artificial intelligence. I also mean research into creative technologies, environments that facilitate creative processes, document systems that facilitate creative insight, and even devices that induce brainwave patterns believed to be correlated with creative activity. These kinds of technologies, if they were applied to the particular case of developing machine consciousness, would be ways of pursuing the bi-directional mode of philosophy/engineering design I have been discussing.

Finally, my own work on the SEER-3 project and "synthetic phenomenology" (Chrisley and Parthemore 2007) is another example of an application of this design strategy. This research aims to produce a robotic system such that the understanding gained by interacting with it permits the specification of experiential states that are not easily specifiable by non-robotic, non-interactive means. Thus SEER-3 is another example of how the skills one acquires by interacting with one's own artefacts might be useful in making some progress on understanding consciousness.

References

Alter, T. (1998). "A Limited Defence of the Knowledge Argument". *Philosophical Studies* **90**, 35-56.

Bickhard, M. (2008). "Interactivism: A Manifesto". Forthcoming in Campbell, R.L., Ó Nualláin, S., & Bickhard, M.H. (Eds.), *The Study of Mind: Toward Inter- and Intra-Disciplinary Cooperation*. Available at http://www.lehigh.edu/~mhb0/InteractivismManifesto.pdf; accessed 11/11/07.

Breazeal, C. and Scassellati, B. (2000). "Infant-like Social Interactions Between a Robot and a Human Caretaker". *Adaptive Behavior* **8**:1.

Chrisley, R. and Parthemore, J. (2007). "Synthetic Phenomenology: Exploiting Embodiment to Specify the Non-Conceptual Content of Visual Experience", *Journal of Consciousness Studies* **14**(7):44-58.

Dennett, D. (2005). *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Cambridge: MIT Press.

Held, R. and Hein, A. (1963). "Movement-produced stimulation in the development of visually guided behavior". *Journal of Comparative and Physiological Psychology* **56**(5):872–876.

Herbert, F. (1966). *Destination: Void*. Penguin.

Hurley, S. (1998). *Consciousness in Action.* Cambridge: Harvard University Press.

Jackson, F. (1982). "Epiphenomenal Qualia". *Philosophical Quarterly* **32**:127-36.

O'Regan, K. & Noë, A. (2001). "A sensorimotor account of vision and visual consciousness". Behavioral and Brain Sciences **24**(5): 883-917.

Wittgenstein, L. (1972). *Philosophical Investigations*. Oxford: Blackwell.