Can a Virtual Entity Support Real Consciousness, and How Might This Lead to Conscious Robots?

Owen Holland

Sackler Centre for Consciousness Science
University of Sussex, Brighton BN1 9QJ, UK
O.E.Holland@sussex.ac.uk

Abstract. A key issue within the area of putatively conscious AI systems is that of whether a wholly virtual system can ever be capable of supporting real consciousness. This paper first considers the theoretical implications if such systems using digital technologies could in fact exist, and then explores the consequent practical implications for the creation of virtual and real conscious systems. It is argued that the best strategy is probably to first create a virtual consciousness with components analogous to a virtual embodiment, and then to migrate the conscious core by degrees into a physical embodiment.

Keywords: Conscious robots, virtual entities, virtual worlds

1 Some Fundamental Thoughts

The key question is easy to ask: Could a virtual entity in a virtual world be conscious? Many people find it easy to answer, too, with a Yes or a No, but this is usually an expression of belief rather than the result of a reasoned review of facts and underlying assumptions. For the moment, let us ignore the arguments for and against the two positions, and instead explore some of the consequences if Yes is the correct answer.

To answer Yes, we would need to have a clear idea of what we meant by virtual. As of now, the only way we have of creating virtual entities and virtual worlds is to use digital substrates – computers in one form or another – and so we will restrict the discussion accordingly. (For practical reasons we omit quantum computers.) In the simplest case, the entities and the world will be represented in terms of the states of many different kinds of physical computer components that have been arranged so that they behave as if they have finite logical states that make synchronised transitions determined only by the previous states and any current inputs. It is important to realise that the phrase 'behave as if' conceals what is really happening at the electromagnetic level. Electronics engineers say that digital is just overdriven analogue, and so continuous change and a lack of strict simultaneity and synchronisation hold sway at the micro and nano scales. This limits any definition of consciousness derived from physical processes.

At the level with which we are concerned, both the virtual entity and its world are also necessarily constructed from these logical states and notionally synchronous

transitions, and their boundaries and causal interactions are defined in the running software, including the operating system, and not at any lower level. However, while the logical state transitions are arranged to be deterministic, transitions at the higher software component level may be made to be effectively probabilistic by specifying probability distributions underpinned by random number generators. A software entity and the software world by which it can be affected, and which it can affect, do not intrinsically have to be representations of anything else. What then separates the entity from its world so that we could say that the entity was conscious, but not its world? This question is potentially complicated by the contention by some that consciousness in nature is not limited to the physical boundary of an entity, but extends into the world [1]. This idea will not be considered further here. However, one thing is certain: if both conscious entity and world are instantiated in the same computer, it would be nonsense to say that the computer is conscious. However, if the world in the computer is derived from and linked to the real world, and if the conscious computer-based entity can cause changes in the entity's world that are reflected in changes in the real world, then it might seem reasonable to many to say that the computer is conscious in exactly the same way as we say that a person is conscious, though this perhaps says more about our imprecise language use concerning humans than anything else.

What are the requirements for certain aspects of running software to constitute an entity that maintains its identity over time? (The concept of a software entity already exists in software engineering, but we mean something rather different here.) In the simplest case, if the memory locations used by the entity specified in the software are static and known at compile time, then this is not a problem, as the boundaries of the entity and the successive processes within it can then be completely known. However, if the locations are dynamic, only known at runtime, and possibly dependent on events, the physical locations used by the entity can vary with time, and so the entity can only be defined at the level of the software structure and not at the hardware level. In addition, there must exist some degree of separation of the entity from whatever else there is in the software – the world, perhaps including other entities – while still allowing each to affect the other in an interaction that must be constrained in certain ways. Perhaps the best strategy is to define the scope of the transactions between them in such a way as to reflect the intrinsic constraints of the physical transactions between our bodies and the world, and this involves assuming that the virtual entity is contained within a structure corresponding to a virtual body. One convenient way of achieving this is to use physics based software to model both the virtual entity's body and its environment. A further advantage of this is that much of what we know about our own consciousness is both shaped and expressed in terms of our physical body and its sensory and physical interactions with the world, and having access to similar representations may help us to appreciate any commonalities between our own consciousness and the putative consciousness of the artificial entity.

We can now extend the argument above that a virtual conscious entity with a virtual body would have to be distinct from its virtual world. Should we go further in limiting the scope of describing something as conscious by restricting consciousness to the entity's equivalent of the human nervous system (assuming it has one)? In normal speech we do not do this in relation to humans – we refer to people as being con-

scious, not their brains, although most scientists accept that consciousness is produced by the brain, or more precisely by parts of the brain. The reason for raising this point here is that the part of a virtual entity that is capable of consciousness could in principle be copied into other virtual or real entities, and this could provide the basis for a technology of conscious agents. However, if consciousness turned out to be inseparable from the virtual body of the entity in which it had first been produced, the basis for a technology would be much less clear.

Even if the correct answer to our original question is Yes, there will still be those who refuse to accept that a virtual entity could be as conscious as we are because it would lack some essential but unknown characteristic linked to physicality. This argument is distinct from arguments such as the one that only systems of biological neurons can support consciousness – it is not the type of physicality that is at issue, but physicality in itself. Proponents of this view would accept the possibility of a physical robot becoming conscious, but not a virtual entity. In this context, it is interesting to consider Metzinger's contention that the phenomenal self is itself a virtual entity existing in a virtual world [2]. If this structure is indeed at the root of consciousness, then a conscious virtual entity in its virtual world would have to possess its own virtual self-model in its own virtual world model – a kind of recursion, and Minsky's exploration of self-modelling should be borne in mind [3]. If correct, this decoupling of our phenomenal self from our own physical substrate could imply that the apparent consciousness of the phenomenal self of a virtual entity in a virtual world would be as real and in some deep sense as valid as that of a physical entity in a physical world.

We argue now, somewhat against ourselves, that the relationship between a putative self model and its associated world model may not be comparable to that between a real or virtual entity and their corresponding real or virtual worlds. A useful example here is the artificial life software game Creatures, written by Steve Grand, and first released in 1996 [4]. Players could rear, educate, and care for artificial creatures that were modelled at the genetic, biochemical, and neural levels, and which could interact with conspecifics and objects in their simulated world. However, the relationship between the creatures and their world was not in terms of highly granular simulated sensory perception and sophisticated learning leading to carefully graded and controlled actions, but rather in terms of something closer to message passing between the creatures and the objects. This was a necessary consequence of the severely limited available computational resources. The messages from an object effectively signalled the type of the object, along with what would nowadays be called its affordances – the actions that could be executed in response to the object by the creature, and their outcomes. A creature would then select a message specifying the selected action, and the software would execute the appropriate script.

Apart from the phenomena of innate releasing mechanisms and fixed action patterns, there is obviously no such message passing between physical organisms and physical objects, and none is intrinsically necessary between virtual organisms and virtual objects. However, if Metzinger's self models and world models exist, the system supporting them can possess information about their possible interactions that in principle could be made available to both self model and world model, potentially

creating something closer to the Creatures concept: a virtual entity's self model interacting with an object in its world model would not have to engage in a perceptual and behavioural learning process to establish and respond to the object's affordances – it could simply be given them by the supporting system, and the object would similarly behave appropriately as a result of any action by the self model. This hypothetical process would be independent of the reality or virtuality of the original entity. There is not the space here to trace any correspondences between the behaviour of these self and world models and the experience of consciousness, but this does not matter: what is important is the possible difference between their behaviour and that of of the base level real or virtual entities and worlds. In passing we note that building the self and world models using physics based modelling is a convenient but potentially limiting way of enabling the support system to provide information about interaction to both self and world models.

2 What is to be done?

Whether we are to create an artificial consciousness in order to understand natural consciousness, or to provide a basis for a technology of consciousness, we have a choice of doing it either in a real robot in the real world, or in a virtual robot in a virtual world (or if experiencing social interaction with conspecifics is necessary to become conscious, many real or virtual robots), or in both. From the point of view of efficiency, tractability, access to information, and sheer practicality, the virtual option beats the real option hands down. Working with real robots and the real world is slow and difficult: robots break, age, wear out, need updating, are expensive, and need constant engineering attention; the real world is incompletely knowable, and also changes with time. Further, it is impossible to speed up real time, to which robots and the real world are unavoidably confined, or to repeat an experiment exactly. With a virtual approach, changes are easy, knowledge is in principle complete, multiple parallel implementations can be run, real time is not a constraint, and the base technology improves at an exponential rate. Further, once artificial consciousness is achieved in a virtual entity, the components supporting consciousness can in principle be connected with, or cloned and inserted into, a real robot, because the only available means for controlling a real robot is again the use of software running on a digital computer.

If we accept all of the foregoing, how should we structure an attempt to produce a conscious physically embodied entity? To some extent this depends on the reasons for wanting to create such an entity: is the primary aim scientific, the understanding of consciousness and in particular human consciousness? or utilitarian, the deployment of conscious physically embodied agents for some practical purpose, such as companion robots that do not deceive the user by merely pretending to be conscious? For the first, the physically embodied entity need only exist in a limited, carefully designed environment, but for the second a more open environment would have to be be targeted. Here we will only consider the first option, and the opportunities it offers.

For arranging a transition between the virtual and the real, the easiest and probably most rewarding initial strategy would be to provide both real and virtual environ-

ments, and real and virtual embodiments, that were as similar as possible. For each of these, the choice of strategy boils down to either creating the virtual component and then building the real component to match, or starting with the real component and modelling it in software to produce a corresponding virtual component. For many reasons, the second option is to be preferred. Firstly, it enables the automatic construction of parts of the virtual models from real data, which is much easier than doing the opposite – in particular, the physical characteristics of objects in the real environment can be used to programme the object parameters in the physics based modelling system. Secondly, it offers the prospect of a gradual migration from the virtual to the real environment by using some real sensory data from the real environment, especially visual data, within the virtual world. Thirdly, the performance of the virtual embodiment can be calibrated by measurements taken from the real embodiment, rather than trying to build a real embodiment accurately matching the virtual embodiment. In this context it is worth noting that many available research robots are now supplied along with their physics based models, and for others third-party models are often available – for example [5].

In relation to consciousness, it is now quite well established (by research beginning with [6]) that a normal individual's identification with his/her actual physical body is quite labile, and that by various technical manipulations the sense of self can be transferred towards or even into other locations or embodiments. This may perhaps enable the progressive substitution of more sophisticated embodiments as the project continues and technology advances, the implication being that after a period of adaptation the old embodiment could be functionally replaced with the new without disturbing the underpinnings of consciousness itself. If posssible, this will also validate part of the potential path for the exploitation of conscious robots, although the move to an open environment may prove a more serious problem. When asked why he insisted on using real robots rather than making things easier by using simulations, Rod Brooks replied that there was certainly no problem in simulating his relatively simple robots – it was simulating the rest of the world that was difficult, and this might prove to be a major problem if the strategy of developing conscious robots in simulation and migrating them to reality turned out to be the only practical option.

3 Conclusion: Learning from an early effort

Several of the ideas discussed above were explored in the CRONOS project, an early attempt [7] to build a robot with some features of consciousness. The central idea, inspired by Craik's work [8] but in many ways similar to Metzinger's [9], was to build a robot with a body similar to a human – an anthropomimetic robot – with a jointed skeleton and compliant musculature, to equip it with an internal model of itself and its immediate surroundings, and to investigate its cognitive abilities in relation to Aleksander's axioms of consciousness [10], Metzinger's constraints [9] and Tononi's phi [11]. It was of course only partially successful, but it provided many opportunities for learning the lessons leading to the above formulation. The robot's skeleton was hand moulded, and the robot was fitted out with a variety of poorly specified commercial

components. This was partly due to a scarcity of funds, and partly deliberate – our biological components are generally of poor quality, with characteristics that are constantly changing, yet the brain manages to control them very well indeed. It proved almost impossible to control, and two European follow-on projects developing the concept only managed to improve the controllability mainly by improving the engineering [12, 13]. We hypothesised that the control strategies developed by the brain might have some relevance for consciousness, and this is still a possibility, but if they exist they remain unknown. If this hypothesis is discarded, using a conventionally engineered and controlled robot is a sensible move. The robot's internal model, SIMNOS, was built using the proprietary physics based software Ageia PhysX [14]. The hand-built and partially unknown nature of the robot made this an impossibly difficult task, with an unsatisfactory outcome. The follow-on projects used more modern and configurable modelling software, but the deficiencies of physics based modelling are still a major problem. However, one pioneering CRONOS strategy, that of using the physics based model as a primary virtual entity, and re-using the physics based model as the primary entity's self model [15] probably offers a key strategy for the eventual development of a conscious robot technology.

References

1. Manzotti, R.: An alternative process view of conscious perception. Journal of Consciousness Studies, 13(6) pp. 45-79. Imprint, Devon (2006)
2. Metzinger, T.: The subjectivity of subjective experience: A representationalist analysis of the first person perspective. In: Metzinger, T. (ed.) The Neural Correlates of Consciousness. MIT Press, Cambridge (2000)
3. Minsky, M.L.: Matter, Mind, and Models. Proc. International Federation of Information Processing Congress, vol. 1, pp. 45-49. (1965)
4. Grand, S.: Creation: Life and how to make it. Harvard University Press, Cambridge (2003)
5. gazebosim.org last accessed 2018/12/12
6. Petkova, V.I., Ehrsson, H.H.: If I Were You: Perceptual Illusion of Body Swapping. PLoS ONE 3(12): e3832 (2008)
7. Holland, O.: A strongly embodied approach to machine consciousness. In: Chrisley, R., Clowes, R., Torrance, S. (eds.) Journal of Consciousness Studies Special Issue on Machine Consciousness. Imprint, Devon (2007)
8. Craik, K.J.W.: The Nature of Explanation. Cambridge University Press, Cambridge (1943)
9. Metzinger, T.: Being No One: The Self-Model Theory of Subjectivity. MIT Press, Cambridge (2003)
10. Aleksander, I., Dunmall, B.: Axioms and tests for the presence of minimal consciousness in agents. Journal of Consciousness Studies, vol 10, pp.: 7-18. Imprint, Devon (2003)
11. Tononi, G.: An information integration theory of consciousness. BMC Neuroscience, 5 (1), pp. 42. (2004)
12. eccerobot.org last accessed 2018/12/12
13. https://cordis.europa.eu/project/rcn/102206_en.html last accessed 2018/12/12
14. https://en.wikipedia.org/wiki/PhysX last accessed 2018/12/12
15. Marques, H.G.: Functional Embodied Imagination. Ph.D. thesis, University of Essex (2010)