Must Machines be Zombies? Internal Simulation as a Mechanism for Machine Consciousness

Germund Hesslow and Dan-Anders Jirenhed

Department of Experimental Medicine, Lund University
BMC F 10, SE 221 84 LUND, Sweden
Germund.Hesslow@med.lu.se
Dan-Anders.Jirenhed@med.lu.se

Abstract

One of the many problems of consciousness concerns the appearance of an inner world or an inner reality. Having an inner world is a prerequisite for consciousness in a machine. In this paper we will argue that the core of an inner world is already present in a very simple robot. We have previously argued that a crucial mechanism for generating an inner world in humans is the ability of our brains to simulate behaviour and perception. A simple robot has been designed in which perception can be simulated. We argue here that this robot, or one that has been extended in various ways, but without adding any new fundamental principle, has an inner world and subjective experience in the same sense as humans.

1. Introduction

Many writers have found it reasonable to suggest that there could be, or that evolution could have produced an organism looking like a human being and responding just like a human to any given input, but without the accompaniment of an inner world of experiences, thoughts and feelings. Such zombies, as they are known, would be indistinguishable from humans. They would behave exactly as we do and would therefore be able to survive and reproduce to the same extent. But if zombies are possible, why did evolution produce a mechanism for generating an inner world? And what is that extra thing, beside our ability to respond to sensory input, that we call the experience of that input?

The question can be rephrased for machines. If we design a robot that can respond to visual stimuli in ways that seem purposeful or adaptive by reasonable standards, would that be an example of a zombie in the philosophical sense? It would clearly be able to see things, but it also have experiences of those things? Or would we have to add some additional mechanism to achieve this? If so, what? Is there, to use the currently popular jargon *something it is like* for the robot to see something?

We have previously argued that the experience of an inner world is generated in biological organisms by a mechanism for internally simulating interaction with the external world Hesslow, 1994; 2002). We have also made some initial attempts to implement this simulation mechanism in a simple robot. (Jirenhed et al., 2001; Ziemke *et al.*, 2005, Hesslow & Jirenhed, 2007). Here, we want to discuss how far the simulation mechanism will take us towards providing some key properties of consciousness for this robot or for a robot that has been extended in various ways. In particular, we want to discuss if such a robot can reasonably be said to have an inner world and to experience things in the same sense as a human being.

We will first briefly explain the proposed simulation mechanism and the basic design of the robot in which this mechanism has been implemented. We then discuss if this robot, henceforth called *K*, can really be said to have an inner world and if it can be said to have experiences.

2. The simulation mechanism

The purpose of designing *K* was to implement a *simulation* mechanism that we believe holds the key to explaining the appearance of an inner world in humans (Hesslow, 1994;Hesslow, 2002). The simulation mechanism, as it applies to human beings, has three components.

a) It assumes that an action can be simulated by activating motor structures in the frontal lobe roughly as they would be activated during an overt action, except that the final motor output is suppressed. Simulated actions are essentially low amplitude behaviours.

b) The second, and in the present context most important, assumption is that perception of an external stimulus can be simulated by internally elicited activation of the sensory cortex in a way that is similar to the way it would have been activated by normal perception of an external stimulus. Thus, if my visual cortex is activated in a way that sufficiently resembles the activity that occurs when I am looking at a real tree, the neural processes that follow will also be similar. Thus, the neural activity that normally underlie seeing a tree will occur, regardless of how that activity is elicited and regardless of whether any tree or tree-like object actually exists. Although seeing a tree would normally entail the existence of a tree that is seen, the simulation process suggests that there is a sense in which we can be said to see a tree even if there is no tree to be seen. Another way of putting it is to say that simulated perception can explain why there *appear* to be such objects in spite of the fact that there are good reasons to deny their existence.

c) Thirdly, we assume that there is an anticipation mechanism such that early stages of both overt and covert actions can elicit perceptual simulation of their normal consequences. Such perceptual simulation will normally occur before the action has been performed and also when the overt performance of a prepared action is interrupted.

A consequence of these mechanisms is that a simulated action can generate simulated sensory activity, which in turn can function as a stimulus for new simulated behaviour and so on. Thinking, on this view, is essentially a simulated interaction with the external world.

Although it is by no means generally accepted, there is a wealth of evidence that such a simulation mechanism exists and it has been argued elsewhere that it can explain many aspects of cognitive function and the appearance of an inner world in biological organisms (Hesslow, 1994;Hesslow, 2002 and in robots (Ziemke et al., 2005; Hesslow & Jirenhed, 2007).

3. The reality of the inner world

The simulation hypothesis suggests that we humans have an inner world in the sense that we can experience a world through simulated perception even though there are no actual objects corresponding to the experienced objects.

There is of course a logical problem here. My seeing something is usually taken to entail that there is something there to be seen and if there is nothing there to see, it follows that I am not seeing. We can choose here between two ways of talking about this. On the one hand, we could say that the inner world is an illusion. It does not really exist, but *simulation explains why it appears that it exists*. On the other hand, we could say that *simulation is the mechanism whereby the inner world is created*.

The idiom we choose here is largely a matter of taste. We will talk here about the inner world as if it was real, but we want to make it clear that it's reality is very different from the reality of the external world and also from the world of mental objects envisioned by many philosophers. It is not assumed that simulation creates any internal images, for instance.

A person, who has lost a limb, can still have a clear perception of it and can have severe pain which seems to emanate from their so called ‘phantom’ limb. The phantom limb is quite real to such a patient, and talking about it as if it really existed is difficult to avoid, but could be seriously misleading if someone was led to believe that the phantom limb was made of flesh and blood.

To take an analogy, if a mechanical object is designed to respond in a certain way to a certain input, it is usually possible to bypass the input and insert a ‘fake’ input before the response-generating mechanism. For instance, an external signal could make the speedometer of a car display a speed when the car was actually standing still. This is also true if the response is a verbal report of the input. My computer can perceive which key is being been pressed on its keyboard and can display a report on the screen ‘I notice you just pressed *Y*’. It would not be difficult to bypass the keyboard and send a similar signal to the central processor, tricking it into displaying the same message. The computer would then be simulating the observation of *Y* being pressed, but we should not be tempted into saying that there is an ‘image’ of Y being pressed or a mental representation of such a press in the CPU.

4. The robot

K is a small robot with a very limited behavioural repertoire (Ziemke *et al.*, 2005). It has two motor driven wheels (one on each side of its body) that can be controlled individually and allow it to move around with forward, backward and turning motions. *K* receives sensory input via a very simple camera, mounted on top of its body, that has an artificial retina with a resolution of 10 pixels.

The robot is confined to a simple environment consisting of four corridors that form a square and at the end of each corridor there is an object that is detectable by the robot's camera. When the robot moves around in the environment, the objects cause retinal activation when they are within the camera's field of vision. The retinal activation pattern depends on the relative angle and distance between the robot and the object. As the robot moves down a corridor, gradually approaching an object, the retinal activation gradually increases in strength and extent.

Within the robot is a controller network – a feed-forward artificial neural network (ANN) that has input neurones activated by the artificial retina in the camera. The controller network is comprised of two modules that produce different outputs, first a behaviour module that controls the motion of the two wheels based on retinal activation, and second, a prediction module that produces a simulated retinal activation pattern based on the current retinal input and also the motor output that is elicited by the behaviour module. The simulated retinal activation pattern that is produced can be used as input to the network, standing in for the input that is normally supplied from the environment.

The strengths and signs (positive or negative) of the connections between neurones in the network are developed in two stages using a genetic algorithm. First, the neuronal connections within the behaviour module are adapted so that the robot acquires the ability to move around in its four-corridor environment without collisions. Second, while the adapted behaviour module remains fixed, the connections within the prediction module are adapted. Since the function of the prediction module is to produce a retinal activation pattern that is similar to that which actually follows as a consequence of the robot's behaviour, one possible training strategy is to reinforce simulated perception that in some way resembles the actual perceptual outcome of the behaviour. This approach was tried in a previous investigation (Jirenhed *et al.*, 2001), with the result of over-generalization. Events that rarely occurred, e.g. corners, could not be predicted, regardless of how important they were for navigation.

A different strategy, one that was chosen here, is to use whatever predicted retinal activation pattern the module produces and evaluate it based on the behaviour it produces via the behaviour module. With this method of indirect reinforcement of good predictions, the robot acquired an ability to control its behaviour based on simulated perception alone for sequences with a length of up to several hundred time steps.

The robot had thus learned to first move around in its environment relying on externally generated perceptions (i.e. retinal activation via the camera) and then to do the same navigation based solely on its internally generated perceptions (i.e. retinal activation via the prediction module).

We have to emphasise that *K* is not actually a physical robot but merely a computer simulation of one. However, the simulation is based on actual sensor inputs and it has previously been shown that results obtained with a similar simulator can be transferred to physical robots (Miglino *et al.*, 2007). We therefore feel justified in talking about K as if it was a physical robot. Notice also that when we speak of simulation below, we always refer to internal perceptual simulation in the robot, not computer simulation of the robot.

5. Does K have an inner world?

Given that K can simulate perception of something it is not really seeing, because the sensor input is being generated by itself, does it have an inner world in the same sense as a human being? We would like to suggest that it does. *K* is not only able to see objects in its environment, but also to simulate such seeing by having predicted retinal inputs taking the place of external inputs and being fed into its internal circuitry. It would seem then, that *K* has the same ability that we have argued lies behind the inner world of humans. If this is right, it must be legitimate to claim that *K* too has can have an experience of something that does not exist in the external world but only, so to speak, in *K*'s 'mind'.

There are several obvious differences between the abilities of humans and those of *K*, and many of these differences could be taken to justify the rejection of our claim.

For instance, it could be said *K* is merely mechanically responding to events in its internal circuitry and that this is true of many mechanical gadgets. After all automatic vacuum cleaners and lawn mowers also have internal electronic circuits that control their behaviour and they are able to navigate more complex environments than *K*. Yet, we do not say that these machines have inner life. What is so special about *K*?

The crucial difference is the way that these machines are controlled. There is clearly a sense in which *K* can 'see' its environment, when the sensors are turned on, just like the vacuum cleaner or lawn mower. But, when they are turned off, *K* is using the internally simulated inputs instead. Thus, there is a part of *K* that receives the same input when it is simulating seeing as when it is actually seeing. From the point of view of this part, then, there is little difference between real and simulated input. The situation is very similar to that when a human being 'feels' a phantom limb and to whom the phantom limb could be said to 'exist', though this is clearly a misleading way of putting it.

The difference between *K* and the other machines is not just a question of memory. It is clearly an important feature of *K* that it can 'see' things when the sensors are turned off. But other machines could be programmed to navigate in an environment using stored information instead of sensor input. Could it not be said that such a machine can also 'see' an environment internally? To a certain extent, the answer is a matter of taste and definition, but we want to stress that the simulated seeing of *K* is in crucial respects similar to actual seeing and not equivalent to any internal mechanism for eliciting the right behaviour.

6. What is *K* seeing?

It may seem somewhat contrived to say *K sees* a particular object at all. There is no difference for *K* between seeing object A and object B and there is really no point in saying that it 'sees' object A if this is the same as seeing B and when all the input does is to elicit a specific avoidance response. This objection is reasonable but can be met by enabling *K* to respond in different ways to different objects.

To begin with, suppose that there are at least two different kinds of objects in the environment and that *K* learns different responses to them. At present, it can only see obstacles and avoid them by turning right, but we could easily have *K* turn right in front of, say, tall objects and left in front of short ones. What *K* can see is actually a reflection of its ability to respond differentially. There is a large redundancy in its visual input but there is no sense in saying that it can 'see' various aspects of these inputs if it does not make any use of them. However, if it responds differently to low and tall objects, it is meaningful to say that *K* can see the difference.

Furthermore, we can add more abilities to respond to the same external stimulus such as pointing to an object in the environment, drawing it or even giving a verbal description of it or. Which of these behaviours actually occurs on a particular occasion could be determined by other stimuli. For instance, *K* could have a microphone and a module that recognized the question: "What are you seeing?". The stimulus combination consisting of this question and object X or Y could elicit the responses "I see an X" or "I see a Y", respectively. Equipping K with verbal behaviour is of course going far, but the point is only to explore the consequence of enriching K's behavioural repertoire. Even if *K's* linguistic abilities were limited to simple mechanical naming of different objects, the impression that its behaviour is controlled by an inner reality would become much stronger.

Since *K* has the ability to simulate internally its perception of the environment, all these behaviours could then be elicited in the absence of any external stimulus. *K* would thus be able to avoid, point to, draw and verbally describe an object that is not physically present. Surely, it would be reasonable to say that the internal stimulus that generates these behaviours can make *K* see a particular object. If it can generate this stimulus internally, it would also be reasonable to say that it can imagine the corresponding object and that an inner reality will exist in or appear to the robot.

We have a problem of mental objects that is quite analogous to that raised by human minds. In both cases, it can be dissolved by the simulation mechanism. When *K* is avoiding, pointing to, describing etc. a simulated perception of an object in its environment, it becomes tempting to inquire about the nature of that to which it is responding. But there is no 'object' to which *K* responds. Although *K's* response to obstacles is elicited by physical processes in *K*, the obstacles themselves are not present in it. Regardless of how we eventually decide to solve it, the logical problem of how should talk about internally perceived things, is not specific to *K*, has nothing to do with its being a robot, and cannot be a reason to reject our proposal that *K* has an inner world.

7. Similarity of imagined and real objects.

A very interesting question concerns the similarity of the simulated and real perceptions. In a previously reported series of robot experiments, we compared the sensor activations caused by external stimuli with the simulated sensor activations (Ziemke et al. 2005). Because the simulated perceptions elicited the same behaviour as the real sensor activations, we expected the activation patterns of the sensors to be similar. Surprisingly, they were quite different. It could be argued that the apparent lack of similarity between externally caused and predicted sensor activations means that the simulated perception is not really a simulated perception at all. This is a potentially serious objection because we usually regard it as an important aspect of the inner world that it resembles the external world, except that it is mental rather than physical. To this we would respond with the following observations:

Firstly, if *K* had been taught to do other things than just avoiding obstacles, such as naming, drawing, pointing to etc. the similarities would probably have been greater. We have not done the actual experiments, but consider an analogy. If obstacles vary in colour but *K* is not required to respond differentially to different colours, perceptual simulation will work just as well regardless of how the colour circuits of its sensors are activated and no similarity in this respect should be expected.

Since K is only required to do one thing in response to its input, there is an enormous redundancy in the sensory information. What particular aspect of this input that K will learn to utilise is largely a matter of chance. Tall or small objects, black or grey, round or square will all serve the purpose of avoiding collisions just as well and which feature K happens to use makes no difference. But these possible variations in what features K can use to avoid obstacles will be sharply reduced if other kinds of responses are required.

Secondly, similarity is a relative concept. Whether two objects should be deemed similar, depends on which features are considered relevant. If you are colour-blind but have a good sense of geometry, a black and white picture will probably look very similar to a colour picture of the same thing. To someone with good colour vision but with a poor understanding of the geometric relations of the object, the pictures will look completely different. An experienced chess player and someone completely ignorant of chess will see different things and make different similarity judgments when confronted with pictures of two chessboards in the middle of the games.

Analogously, a small child draws a human being as a head with legs directly attached to it. We do not know much about what goes on the child's brain when it makes such drawings, but it is a reasonable assumption that the head with legs really does resemble a human being for the child and that when the child learns to draw the trunk, it also learns to 'see' human bodies differently.

The apparent lack of similarity between real and simulated sensor activations can actually be turned into an argument *for* the claim that K has an inner world. One of the most powerful 'intuition pumps´ in Dennett's sense (e.g. Dennett, 1988) is the alleged fact that different people can see the same thing and respond to it in the same way, yet have different subjective experiences. The prime example of such an intuition pump is probably the inverted spectrum thought experiment (my experience of red might be the same as your experience of green, and conversely). This is taken to support the idea that in perceiving an external stimulus, there is an irreducible private component in addition to the external one. Given the different sensor activations of Khepera robots apparently seeing the same things and reacting to them in the same way, it could be claimed that *K* too can have an irreducible component in its perception. Indeed, it could be claimed that the way it is for *K* to perceive *X* is different from what it is like for *K'* to perceive *X*. It follows, of course, that there is something it is like for *K* to perceive *X*.

8. Subjective experience

We have argued that there is a sense in which *K* can see *X* even if there is no *X* there to bee seen, because the 'process of seeing' can occur in the absence of *X*. But is this really 'seeing' in any interesting sense? *K responds* to *X* or a simulated *X*, but is this really seeing? We expect that many readers will be uncomfortable with this assumption and feel that a crucial element is missing, namely the subjective experience of (seeing) *X*.

Behind this objection lies an assumption that humans not only respond when they see something, but that the responding is accompanied by an additional element, the experience of *X* or *X*-like qualia. There is 'something it is like' to see *X* that humans have and that we have not demonstrated in *K*.

Since we do not share this assumption (Dennett 1988), we do not think it necessary to argue that *K* has qualia. On the other hand, if there is a physiological mechanism that explains why there *seems* to be an additional element of subjective experience in humans, it would strengthen our case if a parallel mechanism existed in *K*. It does.

As we have observed repeatedly above, when we can simulate seeing an object *X* in the external world, the process of seeing *X* is so similar that it seems we are actually seeing *X*, from which it may seem to follow that *X* exists. The experience of simulation almost forces us to posit the existence of an internal element that accompanies the external one.

Again, we can compare the simulated object with the external one, for instance in making judgments about their similarity. But, of course, a direct comparison cannot be made between an internal and an external object, it has to be between two objects which can exist in the same dimension. That is, we have to juxtapose the imagined *X* and some inner version of *X*. This would again seem to entail the existence of some inner version of *X*, "what it is like" to see *X*, to employ the fashionable idiom.

Thus, the simulation mechanism encourages a certain way of thinking about what goes on our minds and it feeds the intuition that there exists some sort of inner versions of external objects.

K is clearly far too simple to have 'intuitions' or trains of thought that can be 'encouraged' to go in one direction rather than another. Nevertheless, it does have the ability to simulate seeing and some of the questions that are raised by simulation about the inner world of humans, are also naturally raised about simple robots with this ability. We could ask, for instance, if the simulated perceptions of obstacles in its environment, "imagined" obstacles, are similar to the perceptions of actual objects.

The robot cannot ask this question itself. That would require quite sophisticated extensions of its basic capabilities, but this may not be a crucial objection. Small children are also unable to ask such questions, but surely those who believe that there is "something it is like" to see something, would not exclude children. The reason we think that children have qualia is, we suggest, that the intuition pump of internal simulation applies to them.

The intuition that there is a way things look to *K* might be made stronger if *K* had the ability to make similarity judgments. Suppose, for instance, that *K* could judge the similarity between objects *X* and *Y*, by comparing the sensor activations when the two objects were perceived. If this was possible, it would be a small step to assume that *K* could also compare internal simulations and make judgments about the similarity between imagined cases of *X* and *Y*. If this was done, it would seem very natural to speak of how

One way of deciding, from the outside, if two objects look the same or different for K, would be to look at the sensor activations. If they are the same, the subjective experiences are the same, if not, not. But if the subjective experience can be the same or different, then of course the subjective experience must exist.

The intuition pump works even better if we ask the same questions about two different *K*-type robots. Their sensor activations, as we observed above, will normally be different, even when they look at the same object, because of chance events during learning. But then we could start asking questions analogous to, say the inverted spectrum.

Suppose that *K* has a certain internal activation pattern *p* when looking at *X* and *p'* when looking at *Y* and suppose that *K* has a sister *K'* where the patterns are reversed, so that her sensors show the pattern *p'* when looking at *X* and *p* when looking at *Y*.

We do not really want to enter into a discussion of what *K* or his sister are actually subjectively experiencing in these various situations. Our point is rather that we can ask the same questions, *with the same legitimacy*, about the robots as about human beings.

This is not to deny that there are vast differences between the subjective experience of humans and those of simple robots. Any perception will elicit a host of associations in the same and in different sensory modalities and also various emotional responses. If we see a car, images of other cars may pass quickly by and at the same time we may experience the smell of exhaust fumes and petrol, the exhilaration of driving fast. Such associations are an integral part of perception and their nature is clearly very dependent on our being human beings with human perceptual and emotional capacities as well as human life experiences. A simple robot will not only differ from us in the nature of the associations elicited by a particular perception. Robots as simple as *K* will not have any particular associations at all, except in the sense that a particular action may be elicited.

Nevertheless, there does not seem to be any compelling reason to differentiate between humans and a modestly extended version of *K* with respect to the existence of a subjective quality of experience. (For additional arguments for this conclusion based on robots with sensorimotor knowledge, see Kiverstein, 2007)

9. Conclusions

Simulated seeing virtually forces upon us the (false) idea that, when we imagine seeing something, there must be something there to be seen, a copy, an image, a mental object. And if such inner versions of objects really exist, they can be compared to external objects as well as to each other. Because external objects have properties such as colours, and because they can be compared to the inner copies, the latter must also have such properties. Thus, we are encouraged to think about inner versions of real objects that are similar to them. We are led to believe in qualia.

But simulated perception can also exist in robots and the intuitively plausible steps that encourage us to think that humans have qualia can also be applied to *K*. *K* can also simulate seeing. Whatever it is that *K* is seeing can be compared to external objects. Questions can be raised about the similarities and differences between what *K* is seeing and what *K'* is seeing.

It could be objected that *K* and *K'* are not actually seeing existing objects, and it cannot therefore be meaningful to ask how similar these objects are. This scepticism is certainly warranted, but if it is valid, it is also valid for a human being. Thus, whether we believe in qualia or not, if the mechanism underlying the appearance of mental objects in humans and in robots is the same, *K*'s alleged lack of qualia cannot be a reason for denying consciousness to it.

References

Dennett DC (1988) Quining Qualia. In: *Consciousness in Modern Science* (Marcel A, Bisiach E, eds), Oxford: Oxford University Press.

Hesslow G (1994). Will neuroscience explain consciousness? *Journal of Theoretical Biology* 171: 29-39.

Hesslow G (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences* 6: 242-247.

Hesslow G, Jirenhed D-A (2007). The Inner World of a Simple Robot. .*Journal of Consciousness Studies* 14:85-96.

Jirenhed D-A, Hesslow G, & Ziemke T (2001). Exploring internal simulation of perception in a mobile robot. In *Fourth European Workshop on Advanced Mobile Robots (Eurobot '01)*, eds. Arras, Baerveldt, Balkenius, Burgard, & Siegwart, pp. 107-113. Lund.

Kiverstein, J. (2007). Could a Robot Have a Subjective Point of View? *Journal of Consciousness Studies* 14:127-40.

Miglino O, Lund HH, & Nolfi S (2007). Evolving Mobile Robots in Simulated and Real Environments. *Artificial Life* 2: 417-434.

Ziemke T, Jirenhed D-A, & Hesslow G (2005). Internal Simulation of Perception: A Minimal Neuro-Robotic Model. *Neurocomputing* 28: 85-104.