Susan Blackmore

Consciousness in Meme Machines

Setting aside the problems of recognising consciousness in a machine, this article considers what would be needed for a machine to have human-like consciousness. Human-like consciousness is an illusion; that is, it exists but is not what it appears to be. The illusion that we are a conscious self having a stream of experiences is constructed when memes compete for replication by human hosts. Some memes survive by being promoted as personal beliefs, desires, opinions and possessions, leading to the formation of a memeplex (or selfplex). Any machine capable of imitation would acquire this type of illusion and think it was conscious. Robots that imitated humans would acquire an illusion of self and consciousness just as we do. Robots that imitated each other would develop their own separate languages, cultures and illusions of self. Distributed seflplexes in large networks of machines are also possible. Unanswered questions include what remains of consciousness without memes, and whether artificial meme machines can ever transcend the illusion of self consciousness.

I am going to set aside some of the major problems facing machine consciousness and concentrate on the question of what sort of machines might acquire human-like consciousness.

The main problem to be ignored is that we do not know how to recognise consciousness in a machine. That is, there is no obvious equivalent of the Turing test for consciousness. I shall define consciousness here in terms of subjectivity; what is sometimes known as 'phenomenal consciousness' (Block, 1995) or 'what it's like to be' (Nagel, 1974). With consciousness being subjective, any objective test, such as any variation on the Turing test, fails to grasp it. You could certainly have a test that shows whether other people *think* a machine is conscious but this is not the same thing, as our eager propensity to attribute feelings and intentions to even the simplest of robots and mechanical toys reveals. Once we start asking whether there is *really* something it is like to be the machine, or whether the world appears a certain way *for* that machine, then our usual tests fail.

In fact we don't know how to recognise consciousness in anything at all. As far as other humans are concerned this is the problem of other minds, but we usually ignore it on the grounds that we think we know what our own consciousness is like and we then extrapolate to others. We cannot do this so easily for other

Journal of Consciousness Studies, **10**, No. 4–5, 2003, pp. ??–??

species, hence the problem of animal consciousness, and it is even more difficult with machines.

This first problem is exacerbated by the suspicion that there may be kinds of consciousness utterly different from human consciousness, or indeed from any naturally occurring kind of consciousness. This suggests many interesting lines of thought, but many difficulties too. So I am going to ignore this here, and stick to questions concerning only human-like consciousness. This is quite hard enough to be going on with.

Having put these problems aside I shall ask how we might set about constructing a machine that will have human-like consciousness. I am not a programmer or robot engineer and my purpose is not to discuss the details of construction (I would not be able to do so) but is instead to consider the general principles involved. All theories of consciousness have implications for how consciousness might be artificially created, and pondering these may help us better understand those theories. The theory to be discussed here is that ordinary human consciousness is an illusion created by memes for their own propagation. The implication for machine consciousness is that only a machine capable of imitation would develop a human-like illusion of consciousness.

I shall take the most obvious approach, which is first to ask how the illusion of consciousness comes about in humans, and then use that to ask how it might be artificially created. I shall first make some comments about the general nature of consciousness, then consider how it arose in humans, both during evolution and during individual development, and then see what implications this has for machine consciousness.

Consciousness as Illusion

If we hope (or fear) to make a conscious machine it would be helpful to know what consciousness is. We do not. I shall not claim here to solve the hard problem, or to say what consciousness ultimately is (if anything). Instead I shall argue that ordinary human consciousness is an illusion. Therefore, making a machine that is conscious in the same way as we are means making one that is subject to the same kind of illusion. Before explaining this in more detail I want to distinguish this view from some other major positions on machine consciousness, crudely divided here into three.

1. Machine consciousness is impossible

Among those who have argued that machine consciousness is impossible are dualists, those who believe in a God-given soul as the seat of consciousness, eliminative materialists who do not believe that consciousness exists in the first place, and those who argue that there is something special about biological brains that precludes anything else from having human-like consciousness. This last is particularly confusing but the problems are well known and discussed (Dennett, 1995; McGinn, 1987, pp. 279–88; 1999; Turing, 1950). It is worth noting that Searle, in spite of his theory of biological naturalism (Searle, 1992), does not

exclude the possibility of machine consciousness. Rather, he says that any machine could be conscious if it had the same causal properties as a biological brain (Searle, 1997). Since he does not say what those properties are, this is no help in creating artificial consciousness.

Arguments against computational functionalism and good-old-fashioned Strong AI (e.g., Searle's Chinese Room thought experiment), are considered by some to weigh against the possibility of machine consciousness. A more common view is probably that while syntax (or running a formal program) is not sufficient for semantics, and symbols must somehow be grounded in the real world, this can be done by allowing a machine to interact with the world. Developments in neural networks, embodied cognition and situated robotics (e.g., Brooks, 1991; Clark, 1997) suggest the same thing. So this line of argument does not preclude the possibility of conscious machines. Finally, there may be some people who apply to consciousness, rather than to intelligence, what Turing called the 'Heads in the Sand' objection; 'The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so'. (Turing, 1950). None of these arguments provides a good reason for thinking that machine consciousness is impossible.

2. Find consciousness and put it in a machine

Perhaps there is an 'X', or 'extra ingredient', that if we could give it to machines would ensure they were conscious (Chalmers, 1995). McGinn calls the property that would explain consciousness C*, and asks whether it is possible in inorganic materials or not (McGinn, 1999). Some theories of consciousness can be used to derive an 'X' and so to suggest how it could be given to a machine.

I shall not review all the many theories here, but will take just one example as an illustration; the currently popular Global Workspace Theories (Baars, 1988; Dehaene and Naccache, 2001; Dennett, 2001). GWTs equate the contents of consciousness with the contents of a global workspace, likened to a bright spotlight shining on the stage of working memory (Baars, 1988). The GW is a large network of interconnected neurons. Its contents are conscious by virtue of the fact that they are made globally available, or broadcast, to the rest of the system, which is unconscious. I have argued elsewhere that this notion is incoherent and cannot explain consciousness (Blackmore, 2002). GWT depends on the assumption that at any time there is a valid answer to the question 'what is in consciousness now', and that things can meaningfully be said to be either 'in' or 'out' of consciousness, as though consciousness were a container. This is a version of Cartesian materialism (Dennett, 1991) and cannot, I suggest, explain consciousness.

On these theories, 'X' is having a GW. So presumably a machine should be conscious if it is designed with a GW whose contents are broadcast to the rest of its system (see Franklin, this volume). Unfortunately, as mentioned above, even if such machines were built, it would be impossible to test whether they were conscious or not. I can only say that I do not believe that GWT is the way to understand consciousness, and in any case it will have to be tested by other means than making such machines. In the mean time I prefer the third approach, which is to say that consciousness is not what it appears to be.

3. Consciousness is an illusion.

There are several ways of thinking about consciousness as an illusion. Most important is to distinguish them from the view that consciousness does not exist. To say that consciousness is an illusion is to say that it is not what it appears to be. This follows from the ordinary dictionary definitions of 'illusion', for example, 'Something that deceives or misleads intellectually' (Penguin); 'Perception of something objectively existing in such a way as to cause misinterpretation of its actual nature.' (Webster). This point is frequently misunderstood. For example, Velmans (2000) wrongly categorises Dennett's position as eliminativist when it is better described as the view that consciousness is an illusion. I shall explore a version of this position here.

On this view, human-like consciousness means having a particular kind of illusion. If machines are to have human-like consciousness then they must be subject to this same kind of illusion. I shall therefore explore one theory of how this illusion comes about in humans and how it might be created in machines; the theory of memetics.

Human Beings as Meme Machines

Memes are ideas, habits, skills, stories or any kind of behaviour or information that is copied from person to person by imitation (Dawkins, 1976). They range from single words and simple actions to the vast memeplexes (co-adapted meme complexes) of science, art, religion, politics and finance. There are interesting difficulties concerning definitions (Aunger, 2000; Blackmore, 1998), and whether memes can be said to be replicated by means other than imitation, but the essential point is this. When people copy actions or words, those actions or words are copied with variation and then selectively retained and copied again. In other words the actions and words (the memes) fulfil the conditions for being a replicator in a Darwinian evolutionary process (Dawkins, 1976; Dennett, 1995).

This new evolutionary process can only run if the replication process is good enough (has high enough fidelity). Some species of birds, and some cetaceans, can copy sounds with high fidelity, and their songs are therefore memes. But very few other species can imitate at all. Even chimpanzees and orang-utans are, at best, poor imitators and there is much debate over the extent to which they are really able to copy observed behaviours (Dautenhahn and Nehaniv, 2002). Humans appear to be the only species that readily and easily imitate a wide variety of sounds and actions. This suggests that we alone are supporting this second evolutionary process; cultural or memetic evolution. If this is so, human evolution must have taken a very different course from that of other species once we became capable of imitation. I have suggested that human brains and minds were designed by the replicator power of this new process and that this explains why humans are so different from other species (Blackmore, 1999).

There are two aspects of this that are relevant to machine consciousness. First there is how we living humans got to have such large and peculiarly capable brains (the co-evolutionary story). Second is how our individual minds and our

sense of self and consciousness are designed by memetic pressures (the developmental story). Both are relevant to the possibility of machine consciousness.

Meme gene co-evolution

The human brain is excessively large by ape standards, and has been extensively redesigned for language (Deacon, 1997). There is no generally accepted theory to explain this but all existing theories have in common the assumption that the ultimate beneficiary is the genes, and that the large brain and capacity for language must have been adaptive from the genes point of view (Deacon, 1997; Donald, 1991; Dunbar, 1996; Wills, 1993). I have argued, instead, that both were designed by and for the memes in a process called memetic drive.

Once imitation of sufficiently high fidelity occurs, memetic drive works as follows. In a given population of people, memes compete to be copied. People who are especially good at imitation gain a survival advantage by being able to copy the currently most useful memes. Assuming that imitation is a difficult skill requiring extra brain power, this gives an advantage to genes for bigger brains and better imitation. Increasing imitation then provides scope for more competing memes to appear (both useful and harmful ones), and hence there is pressure to be a selective imitator. One effective strategy might be to copy the 'meme founts' — those skilful imitators who pick up currently popular memes and create new ones from the old. Meme founts acquire both status and better opportunities for mating. They pass on the genes that made them good at propagating *those particular memes*. Memetic drive creates not only bigger brains but brains that are *better adapted to copying the memes that were successful during the previous memetic competition* — whether or not those memes were directly beneficial to people or their genes.

Music and religion are examples. Once people can copy simple sounds, such as humming or drumming, the sounds themselves compete to be copied. People who are best at copying the winning sounds acquire status and a mating advantage. In this way the successful sounds give an advantage to genes for the ability to copy those particular sounds. Similarly with religious behaviours such as rituals and devotions, the winning memes drive brains to become better at imitating *those particular behaviours*. The result is brains that are musical and inclined to religious behaviour.

I have argued that this same process can explain the evolution of language. In general, successful replicators are those with high fidelity, longevity and fecundity. Digitisation of sounds into words may increase fidelity, combining words into novel combinations may improve fecundity, and every improvement leads to increased memetic competition. The people who can best copy the winning sounds have an advantage and pass on the genes that gave them that ability. Gradually, human brains would be driven by the emerging language itself. In most theories of language evolution, the ultimate function of language is to benefit genes. On this theory it is to benefit memes.

Underlying these examples is the general principle that replicators co-evolve with their replication machinery, just as genes must once have co-evolved with

their cellular copying machinery. In the case of human evolution, memetic evolution drove the genes to construct better meme-spreading brains. More recent examples include the invention of ever better meme spreading devices from roads and railways to the telegraph, telephone and email. In each case the products copied helped spread the copying machinery which in turn made more products possible and so on. From the memes' point of view the internet is an obvious step in improving meme-copying facilities. As Ridley points out 'memes need a medium to replicate in. Human society works quite well; the Internet works even better' (Ridley, 2003, p. 222). It is in this context that I want to look at the possible development of conscious machines.

Mind design by memes

The second relevant issue is how memes design individual minds; that is, how the design process of evolution unfolds in the case of individual people infected with a lifetime of competing memes.

We spend our lives bombarded by written, spoken and other memes. Most of these are ignored. Some are remembered but not passed on. Others are both remembered and passed on. Some are recombined in novel ways with others to produce new memes. Note that there is much dispute about whether we should use the word 'meme' to apply only to the behaviours themselves, only to the patterns of neural representation (or whatever underlies their storage inside brains), or to both (Aunger, 2000). I shall stick to Dawkins's original definition here, treating memes as 'that which is imitated', or 'that which is copied'. So I shall not distinguish between memes instantiated in books, computers, ephemeral behaviours or human brains, since all can potentially be replicated.

On the memetic hypothesis, human development is a process of being loaded with, or infected by, large numbers of memes. As Dennett (1995) puts it 'Thousands of memes, mostly borne by language, but also by wordless "images" and other data structures, take up residence in an individual brain, shaping its tendencies and thereby turning it into a mind' (Dennett, 1991, p. 254). Language is the mainstay of this process. We have brains specially designed to absorb the language we hear (see above), to deal with grammar, and to imitate the particular sounds of the language(s) we grew up with. By the age of about three years the word 'I' is used frequently and with increasing sophistication. The word 'I' is initially essential to distinguish one physical person from another, but very rapidly becomes used to say things like 'I think', 'I like', 'I want', 'I believe', 'That's mine' and so forth, as though there were a central self who *has* opinions, desires and possessions. In this way, I suggest, a false notion of self is constructed.

There have been very many theories of the formation of this illusory self (Gallagher and Shear, 1999). The difference between other theories and the memetic theory proposed here lies in the question 'Who benefits?'. Previous theories suggest that either the individual person or their genes are the primary beneficiaries; memetic theory suggests that the memes are (Dennett, 1995). I have argued as follows (Blackmore, 1999); once a child is able to talk about his or her self then many other memes can obtain a replication advantage by tagging onto

this growing memeplex. For example, saying a sentence such as 'I believe x' is more likely to get 'x' replicated than simply saying 'x'. Memes that can become *my* desires, *my* beliefs, *my* preferences, *my* ideas and so on are more likely to be talked about by this physical body, and therefore stand a better chance of replication. The result is the construction of an increasingly elaborate memetic self. In other words, the self is a vast memeplex; the selfplex (Blackmore, 1999).

The selfplex and the illusion of consciousness

The result of the memetic process described above is that physical, speaking, human bodies use the word 'I' to stand for many different things; a particular physical body; something inhabiting, controlling and owning this body; something that has beliefs, opinions and desires; something that makes decisions; and a subject of experience. This is, I suggest, a whole concatenation of mistakes resulting in the false idea of a persisting conscious self.

The view proposed here has much in common with James's (1890) idea of the appropriating self, and with Dennett's (1991) 'centre of narrative gravity'. There are two main differences from Dennett. First, Dennett refers to the self as a 'benign user illusion', whereas I have argued that it is malign; being the cause of much greed, fear, disappointment and other forms of human suffering (Blackmore, 2000). Second (and more relevant here) Dennett says 'Human consciousness is *itself* a huge complex of memes. . .' (Dennett, 1991, p. 210).

There is reason to question this. Dennett's statement implies that if a person were without memes they would not be conscious. We cannot, of course, strip someone of all their memes without destroying their personhood, but we can temporarily quieten the memes' effects. Meditation and mindfulness can be thought of as meme-weeding techniques, designed to let go of words, logical thoughts and other memetic constructs and leave only immediate sensory experience. The nature of this experience changes dramatically with practice, and it is common for the sense of a self who is having the experiences to disappear. This same selflessness, or union of self and world, is frequently reported in mystical experiences. But far from consciousness ceasing, it is usually described as enhanced or deepened, and with a loss of duality. *If* this experience can justifiably be thought of as consciousness without memes, then there is something left when the memes are gone and Dennett is wrong that consciousness *is* the memes. It might then be better to say that the ordinary human illusion of consciousness is a 'complex of memes' but that there are other kinds of consciousness.

This is, however, a big 'if', and raises all the problems associated with first-person exploration of consciousness (see Pickering, 1997; Varela and Shear, 1999). At present we should not think of this so much as evidence against Dennett's view as a motivation for further research and self-exploration. It might turn out that if meditation is even more deeply pursued and the selfplex is completely dismantled, then all consciousness does cease and Dennett is correct.

The alternative I want to defend here is that memes distort consciousness into an illusion rather than constitute it. On this view the underlying consciousness itself remains unexplained but we can understand the particular nature of ordinary

human consciousness in terms of the selfplex. By creating the illusion of self for their own survival and replication, memes are responsible for our false sense that there is always an 'I' having experiences, and for the inherent dualism that bedevils all our attempts to understand consciousness.

On this view many kinds of machine might be conscious, but only a particular kind of machine could be conscious in a human-like, illusory way. It would have to be capable of imitation (otherwise it could not replicate memes) and live in a community of similar meme-sharing machines (otherwise there would be no pressure for memeplexes to form). Such a machine would, if this theory is correct, be a victim of the same illusions of consciousness as we humans are. That is, it would think it had an inner self who was conscious. Ultimately, it would start wondering what consciousness was and trying to solve the hard problem.

We may now consider some implications of this theory for actual machines.

Artificial Meme Machines

There are two kinds of artificial meme machine to consider; those which imitate each other and those which imitate humans.

Machines imitating humans

The strong prediction for machines that imitate humans is that they would come to be conscious in exactly the same way as we do. That is, they would acquire language, social conventions of behaviour, and knowledge of the intellectual world in roughly the same way as we do. In the process of acquiring all these memes, many memeplexes would form, including a selfplex and so they would end up as deluded as we are.

This simple picture is, however, rendered almost completely unrealistic by the fact that humans are not general-purpose copying machines but highly selective imitation devices (Blackmore, 2001). In particular we have complex perceptual, social and communication systems derived from our primate origins (Matsuzawa, 2001), categorisation systems biased towards discriminating certain objects and actions rather than others (Pinker, 1997), dedicated language systems (such as a language acquisition device, innate grammar or language instinct, (e.g., Pinker, 1994), and (if the above theory is correct) specialised capacities to imitate certain behaviours such as music, dance, ritual, art and story telling, rather than others (Blackmore, 1999).

I can only speculate that it is not necessary to construct a machine with exactly these abilities to get the same memetic effect. How similar the abilities must be remains an open question. An analogous question has been faced by those working on the Cog project in trying to design appropriate sensory systems (Brooks *et al.*, n.d.). They have chosen to copy some aspects of human perceptual systems but not all, and presumably this project will discover what differences this makes. However, Cog barely imitates. Cog can point to objects, recognise joint attention and imitate simple behaviours such as head nods, but it was not designed as a meme machine. The position is complicated by the fact that observers often

attribute imitation and other abilities to robots, especially 'sociable robots' such as Kismet, even when they do not have them (Breazeal, 2001). If the theory discussed here is correct, then Cog, Kismet and other robots will never acquire a human-like illusion of consciousness unless they are dramatically redesigned to be capable of true imitation learning.

If they were designed that way, or if future robots are capable of imitating human behaviour, then these robots ought to start taking part in human memetic evolution. If they have identical abilities to those of natural humans then the situation will be equivalent to having more people sustaining the evolutionary process. The more interesting (and probably more likely) possibility is that they are sufficiently like us to join in our culture, but sufficiently different to change it. For example (speculating wildly), let us imagine they have faster processing, far larger storage capacity and instant web access; but poorer sensory systems, less agile body movements and less subtle emotions. In this case, the memes they most enjoy acquiring and passing on will differ somewhat from our favourites of gossip, food and sex, but the robots will think they are conscious because they too will go through the process of acquiring human language, using the word 'I', attributing beliefs, desires and possessions to it, and thereby forming a selfplex.

In addition, we humans would probably find ourselves in friendships and other relationships with these creatures. This would affect the entire process of memetic evolution and transform our culture. In the process we would be changed too but, presumably, we would still think we are conscious and so would they. Both kinds of creature would live with the illusion of being a self experiencing stream of consciousness.

At some point these machines will start wondering whether we humans are *really* like them or not. They might propose a 'reverse Turing test' to see whether we are intelligent, but we would certainly fail that. More relevant here, they might try to invent a 'reverse Turing test' to find out whether we are conscious. They would no doubt confront all the familiar problems of subjectivity and other minds. But by the time this happened we would probably already be treating them as conscious beings like ourselves, whether or not we have resolved the problems of consciousness.

Machines imitating each other

One of the most important predictions of the memetic theory of language evolution is that all you need for language to arise is a community of imitating creatures, or meme machines. This is quite different from theories that assume that the capacity for symbolic representation is needed first, such as Donald's (1991) theory of mimesis (note that mimesis is *not* imitation) and Deacon's (1997) crossing of the 'symbolic threshold'. On the memetic theory, the only evolutionary 'turning point' or 'threshold' to be crossed is the ability to imitate. Once imitation is in place, language evolves and reference emerges naturally as the language develops.

I previously suggested that this could be tested by creating a group of imitating robots (copybots) capable of copying each others' sounds (Blackmore, 1999). It turns out that I overlooked the need for shared gaze in creating shared meaning,

and wrongly assumed that real copybots could not be built for some time. In fact, they have already been built and the prediction is confirmed.

Steels (2000) describes language as a living system that is continuously evolving and adapting, and has modelled its origin in groups of autonomous distributed agents, using both simulations and robots interacting in the real world. In a series of experiments, two robots can detect and make sounds. They play an imitation game in which one produces a random sound from its repertoire and the other recognises that sound in terms of its own repertoire and then reproduces it. The first then tries to recognise it again. In similar experiments, De Boer (1997) showed that in such a system the culturally successful sounds propagate, and shared vowel systems emerge through self-organisation. Steels has extended this work to demonstrate how shared meanings can also arise using robots that have simple vision systems, categorisation systems and lexicons, and that look at simple coloured shapes presented in front of them. Together the robots come to agree on sounds that refer to something about the shapes they are looking at. Other experiments investigate the emergence of syntactic structures and grammar (Batali, 1998; Kirby, 2001; Steels, 1998). The general conclusion from this, and related work, is that language can be treated as an evolving system in which both syntax and semantics emerge spontaneously.

This research is in its early stages but a fascinating question is whether such robots will spontaneously invent self reference. I expect them to do so, and to re-enact the whole process of selfplex creation. That is, they will initially use the word for 'I' to refer to a physical robot, then begin to talk about it having beliefs, desires, possessions and so on, and this will in turn provide the opportunity for memes to cluster around the growing self concept. Once this happens these robots will, like the human-imitating robots above, acquire the illusion that they are a conscious self experiencing an objective world.

Note that it will be difficult for us to understand their language. If they were designed to have perceptual and categorisation systems just like ours then their language would presumably have a similar structure to human languages. In this case we would be able to learn their language just as we learn human languages, by immersing ourselves in their world (unless they provided us with textbooks and dictionaries to study). But if their perceptual and categorisation systems were different from ours then their language might be extremely difficult for us to learn. We would find it hard to understand what they were talking about because they would parse the world in different ways and talk about completely different things from us. For this reason we would be less likely to attribute consciousness to them, whether or not they had invented self-reference, spawned selfplexes, and hence suffered the illusion of being a conscious self.

This world would contain two (or more) completely different kinds of language. Whether this would be a stable mixture I do not know.

Future Machines

Neither of the types of machine described above is likely to be built in large numbers for the following reason. Imitation is an extremely crude and low-fidelity

method of copying information. It is the best that evolution has come up with for copying between individuals but is easy to improve on with digital technology. For example, text and images are now routinely copied with extremely high fidelity all over the world between relatively crude computers linked by the internet. If we want effective robots it will be better to provide them with the capacity to get what information they need this way than for them to go through the slow and poor-quality process of behaviour imitation or actual speech. Also, skills learned by one could be directly transferred to another. Will such machines be conscious in a human-like way? That depends, according to the theory I have outlined here, on whether the machines' memes can gain a replication advantage by being associated with a fictional self. If so a selfplex will form.

One reason for making machines capable of imitation (apart from research purposes) would be to make them behave more like humans. This might be useful for some purposes, but they are more likely to be hybrid, imitation, digital-copying machines than simple imitating meme machines like us. In other words, they would acquire most of what they know by downloading it directly. If humans are also provided with memory chips, implanted additional processing capacity, and direct links to the Web, then we might all end up as such hybrids, as some people have suggested we will (Kurzweil, 1999). Assuming that the explosive increase in memes carries on then there will be intense memetic competition to get copied by these hybrid creatures and therefore pressure for selfplexes to form. Even such highly advanced creatures would fall for the illusion that they have an inner self who is conscious.

Finally, rather than thinking in terms of how 'we' should design future machines, it may be better to think in terms of how the memetic co-evolutionary process will design them. As machines copy more and more information from one to another, the machinery they use will co-evolve with that information, ultimately becoming self-replicating. We are not likely to have control over this process. In the discussion so far, I assumed that one selfplex was linked with one physical machine, as it is with most human beings. There is no reason why this need always be so. Not only might there be the equivalent of multiple personalities within machines, but selfplexes might have no permanent physical home, being purely informational structures using whatever processing resources they can (Kurzweil, 1999). So long as there is competition to be copied by such entities, and the entities have boundaries that repel some memes and accept others, then selfplexes will form. Their illusions will be a little different from ours since they will not believe they inhabit or own or control a body, but they may still think that they are a conscious self having a stream of experiences.

Whether any of these machines will be able to transcend the illusion of consciousness remains a question for another time.

References

Aunger, R.A. (ed. 2000), *Darwinizing Culture: The Status of Memetics as a Science* (Oxford: OUP).
Baars, B.J. (1988), *A Cognitive Theory of Consciousness* (Cambridge: Cambridge University Press).
Batali, J. (1998), 'Computational simulations of the emergence of grammar', in *Approaches to the Evolution of Language: Social and Cognitive Bases*, ed. J.R. Hurford *et al.* (Cambridge: CUP).

Blackmore, S.J. (1998), 'Imitation and the definition of a meme', *Journal of Memetics: Evolutionary Models of Information Transmission*, **2**. http://www.cpm.mmu.ac.uk/jom-emit/1998/vol2/blackmore_s.html
Blackmore, S.J. (1999), *The Meme Machine* (Oxford: Oxford University Press).
Blackmore, S.J. (2000), 'Memes and the malign user illusion', Association for the Scientific Study of Consciousness Conference 'Unity and Dissociation', Brussels, July 2000.
Blackmore, S. (2001), 'Evolution and memes: The human brain as a selective imitation device', *Cybernetics and Systems*, **32**, pp. 225–55.
Blackmore, S.J. (2002), 'There is no stream of consciousness', *Journal of Consciousness Studies*, **9** (5–6), pp. 17–28.
Block, N. (1995), 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences*, **18**, pp. 227–87.
Breazeal, C.L. (2001), *Designing Sociable Robots* (Cambridge, MA: MIT Press).
Brooks, R.A. (1991), 'Intelligence without representation', *Artificial Intelligence*, **47**, pp. 139–59 (also reprinted, with extra material, in Haugeland, ed. 1997, pp. 395–420).
Brooks, R.A., Breazeal, C., Marjanovic, M. Scassellati, B. and Williamson, M.M. (no date), 'The Cog project: Building a humanoid robot', http://www.ai.mit.edu/projects/cog/
Chalmers, D.J. (1995), 'Facing up to the problem of consciousness', *Journal of Consciousness Studies*, **2** (3), pp. 200–19.
Clark, A. (1997), *Being There: Putting brain, body, and world together again* (Cambridge, MA: MIT Press).
Dautenhahn, K. and Nehaniv, C.L. (ed. 2002), *Imitation in Animals and Artifacts* (Complex Adaptive Systems Series; Cambridge, MA: MIT Press).
Dawkins, R. (1976), *The Selfish Gene* (Oxford: Oxford University Press; new edition with additional material, 1989).
Deacon, T. (1997), *The Symbolic Species: The Co-evolution of Language and the Human Brain* (London: Penguin).
De Boer, B. (1997), 'Self-organisation in vowel systems through imitation', in *Proceedings of the Fourth European Conference on Artificial Life*, ed. P. Husbands and I. Harvey (Cambridge MA).
Dehaene, S. and Naccache, L. (2001), 'Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework', *Cognition*, **79**, pp. 1–37.
Dennett, D.C. (1991), *Consciousness Explained* (Boston and London: Little, Brown and Co.).
Dennett, D.C. (1995), *Darwin's Dangerous Idea* (London: Penguin).
Dennett, D.C (2001), 'Are we explaining consciousness yet?', *Cognition*, **79**, pp. 221–37
Donald, M. (1991), *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition* (Cambridge, MA: Harvard University Press).
Dunbar, R. (1996), *Grooming, Gossip and the Evolution of Language* (London: Faber & Faber).
Gallagher, S. and Shear, J. (ed. 1999), *Models of the Self* (Thorverton, Devon: Imprint Academic).
James,W. (1890), *The Principles of Psychology* (London: MacMillan).
Kirby, S. (2001), 'Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity', *IEEE Transactions on Evolutionary Computation*, **5**, pp. 102–10
Kurzweil, R. (1999), *The Age of Spiritual Machines: How we will live, work and think in the new age of intelligent machines* (New York and London: Texere).
Matsuzawa, T. (ed. 2001), *Primate Origins of Human Cognition and Behavior* (Tokyo: Springer).
McGinn, C. (1987), 'Could a machine be conscious?', in *Mindwaves*, ed. C. Blakemore and S. Greenfield (Oxford: Blackwell).
McGinn, C. (1999), *The Mysterious Flame: Conscious Minds in a Material World* (New York: Basic Books).
Nagel,T. (1974), 'What is it like to be a bat?', *Philosophical Review*, **83**, pp. 435–50
Pickering, J. (ed. 1997), *The Authority of Experience: Essays on Buddhism and Psychology* (London: Curzon Press).
Pinker, S. (1994), *The Language Instinct* (New York: Morrow).
Pinker, S. (1997), *How the Mind Works* (London: Penguin).
Ridley, M. (2003), *Nature via Nurture: Genes, Experience and What Makes us Human* (London: Fourth Estate).
Searle, J. (1992), *The Rediscovery of the Mind* (Cambridge, MA: MIT Press).
Searle, J. (1997), *The Mystery of Consciousness* (New York: New York Review of Books).
Steels, L. (1998), 'The origins of syntax in visually grounded agents', *Artificial Intelligence*, **103**, pp. 1–24
Steels, L. (2000), 'Language as a complex adaptive system', in *Lecture Notes in Computer Science. Parallel Problem Solving from Nature* — PPSN-VI. Volume Eds. Schoenauer *et al*. (Berlin: Springer-Verlag).
Turing, A. (1950), 'Computing machinery and intelligence', *Mind*, **59**, pp. 433–60.
Varela, F.J. and Shear, J. (1999). *The view from within: First person approaches to the study of consciousness* (Thorverton, Devon: Imprint Academic).
Velmans, M. (2000), *Understanding Consciousness* (London and Philadelphia: Routledge).
Wills, C. (1993), *The Runaway Brain: The Evolution of Human Uniqueness* (New York: Basic Books).