Meme Machines and Consciousness

Susan J. Blackmore

Department of Psychology, St Matthias College, University of the West of England, Bristol BS16 2JP, U.K.

ABSTRACT

What would it take for an artificial system to have something like human consciousness? Central to human consciousness is *subjectivity* and the notion of an experiencing self. It is argued that the self is more like a story or myth than a persisting entity that *has* free will and consciousness. Memes are information passed from one person to another by imitation, and the theory is proposed that the idea of a persisting self is a memeplex; a group of memes that propagate together. This selfplex is the origin of ordinary human consciousness. Therefore, only systems capable of imitation (and hence of sustaining memetic evolution) can have this kind of consciousness. Two kinds of artificial system are imagined, one that can imitate humans and one that can imitate other systems similar to itself. The first would come to have human-like consciousness. The second might be conscious in a completely novel way.

KEY WORDS: consciousness, artificial consciousness, self, free will, imitation

I wish to ask a common enough question. What would it take for an artificial system to have something like human consciousness? My answer, however, is not so common. It is this—the system would have to be capable of imitation. In other words it must be a meme machine.

I will first explain what I mean by human consciousness, then review the basics of memetics and what it means to be a meme machine, and finally consider what kind of machine might be conscious.

1. CONSCIOUSNESS

The whole topic of consciousness is fraught with difficulties, most of which I shall try to avoid. I want only to make clear the sense in which I am using the word. In 1974 Thomas Nagel asked his famous question "What is it like to be a bat?" (though he was not the first to do so). He thereby focused the problem of consciousness onto the question of subjectivity, or private experience. The bat's consciousness is what it is like *for that bat*, something that neither we, nor even another bat can know for sure. Similarly my consciousness now is what it is like to be me, now. It is the *experience* of sitting here at my desk with the sun shining in, wondering what to type next.

The subjectivity, or privacy, of consciousness is what causes all the problems. We cannot study consciousness in the way that we study anything else, since everyone else's experience is denied us. We may choose, if we wish, to ignore consciousness, as, for example, the behaviorists did for many decades. Yet it is difficult to deny it altogether. The quality of my experience is evident *to me*. I *care* whether I feel happy or not, whether the view is beautiful, and whether I can concentrate hard enough to write well. We may argue eloquently about neural processes, and yet this does not seem to get directly at this quality of being aware.

If we take consciousness seriously, we admit that there is a problem, but what kind of problem? According to Chalmers (1996) the easy questions about consciousness are ones concerning how perception works, how discriminations are made, how thinking operates and so on, but these do not address the "hard problem" of subjectivity. Why is all this processing accompanied by a subjective inner life? Why are physical systems like brains also experiencers? Why is there *something it is like* to be at all?

Chalmers implies, by these questions, that there might well be systems (machines for example) that could carry out perception, discrimination and thinking, as we do, and yet have no concomitant inner life. They would be zombies—creatures that act like us, perhaps even appear indistinguishable from us, but are not conscious.

Such zombies are impossible, say others. Any creatures that could do everything we do would, necessarily, have an inner life like ours. It just comes with the territory. Being capable of thinking and speaking as we do

simply does entail being aware. This argument is put clearly by Dennett (1991) who asks us to imagine a zimbo. A zimbo is a zombie that is capable of self-monitoring, or recursive self-representation. It has higher-order informational states that are about its lower-order informational states. A zimbo like this would fare well in a Turing test, says Dennett. It would describe inner processes, such as thinking and internal imagery, just as you might do. In the process, it would not only convince you that it is conscious, but may equally well convince itself. Such a zimbo would live under the illusion that it had an inner mental screen on which pictures were projected, and an inner mental life. It would be a victim of the benign user illusion of its own virtual machine. We are all zombies like this declares Dennett (p 406). Our consciousness is nothing more than this kind of reflexive self-representation. There is no extra special something called consciousness that may, or may not, be present as well.

On this view there is no hard problem. There are only "easy" (though actually not that easy) problems. When we solve all these, and really understand how the brain works, there will be no more mystery. As Churchland puts it, consciousness "may have gone the way of 'caloric fluid' or 'vital spirit'" (Churchland, 1988, p 301).

Those who believe in the hard problem tend to be (though are not necessarily) dualists—believing that matter and mind are quite different and require different kinds of explanation. Chalmers himself ends up with a dualist interpretation (a version of property dualism that he calls 'naturalistic dualism'). Others try to use quantum mechanics to explain consciousness. For example Penrose and Hameroff (1996) invoke quantum coherence in the microtubules to explain consciousness—an explanation that Churchland (1998, p 121) likens to "pixie dust in the synapses". Those who do not believe there is a hard problem just get on with the complex task of trying to understand the brain better.

I mention these big controversies in consciousness studies, not to claim I can resolve them, but to make it quite clear firstly what I mean by consciousness and secondly how difficult the problem is. By consciousness I mean subjectivity—or *what it's like.* The problem of human consciousness is to understand how it comes about that I have experiences. It seems to me (and I suppose to you) that there is a world out there and a me in here, and I am

experiencing the world. This distinction between the world and the observer lies at the heart of the problem—the "hard problem" and the "mind-body problem"—and is the reason why dualism seems so tempting even though most philosophers believe it is unworkable. So consciousness is intimately bound up with the notion of self. This is what we have to explain and it is extremely difficult.

I am going to suggest that memetics can help. So first of all I must provide a brief overview.

2. WHAT IS A MEME?

The concept of the meme derives from Dawkins's 1976 book, *The Selfish Gene,* in which he popularized the growing view in biology that evolution by natural selection proceeds not for the good of the species or the group, nor even for the individual, but for the genes. Although selection takes place largely at the individual level, the genes are the replicators. They are what is copied, and it is their competition that drives the evolution of biological design.

In explaining this, Dawkins (1976) was concerned not just with genes but with the principle of Universal Darwinism. Darwin's (1853) fundamental insight was this—if living things vary in ways that affect their fitness for survival, and if they produce more offspring than can possibly survive, then the survivors in each generation will be the most fit. And if living things pass on their characteristics to the next generation, then the characteristics that helped those ones survive will be more common in the next generation. This, as Darwin saw, is an inevitable process that simply must occur if the conditions are fulfilled. Dennett has called this the evolutionary algorithm. If you have variation, heredity, and selection, then you must get evolution. You get "Design out of Chaos without the aid of Mind" (Dennett, 1995, p 50).

The conditions for this algorithm are certainly fulfilled for biological creatures, but are there any other replicators on this planet? Examples include the immune system and neural processing, among others (Calvin, 1996; Edelman, 1989; Plotkin, 1993), but Dawkins argued that staring us in the face, though still drifting clumsily about in its primeval soup of culture, is

another replicator—a unit of imitation. He gave it the name "meme" (to rhyme with "cream" or "seem"), from the Greek for 'that which is imitated'. As examples he suggested "tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches." (Dawkins, 1976, 192)

Everything you have learned by copying it from someone else is a meme; every word in your language, every catch-phrase or saying. Every story you have ever heard, and every song you know, is a meme. The fact that you drive on the left (or perhaps the right), that you drink lager, think sun-dried tomatoes are passé, and wear jeans and a T-shirt to work are memes. The style of your house and your bicycle, the design of the roads in your city and the color of the buses—all these are memes. Our culture is swept increasingly often by new memes, as communications and copying facilities improve. Diana's death, Clinton's sexual exploits, Tamagotchis and yo-yos have all passed, like infections, around the planet.

We can see that much of culture consists of memes. However, it is easy to get carried away and think of all knowledge, or all experiences, as memes and this is not helpful. We need instead to stick to a clear definition (Blackmore, 1998). The Oxford English Dictionary defines a meme as follows: "meme (mi:m), *n. Biol.* (shortened from *mimeme* ... that which is imitated, after GENE *n.*) "An element of a culture that may be considered to be passed on by non-genetic means, esp. imitation".

It is tempting to say that the meme is *information* copied from one person to another. This is not misleading as long as we accept that memes cannot be pinned down to Shannon information, nor to any other easily definable kind of information. It means only that *something* has been copied—whether it is a bodily movement, an utterance, a design, or a scientific theory. Whatever it is that is copied—that is the meme. Problems undoubtedly remain here, but we do not need to solve them all before beginning to build a science of memetics.

Dawkins referred to memes as "leaping from brain to brain via a process which, in the broad sense, can be called imitation" (Dawkins, 1976, 192). I am using imitation in this same broad sense.

Using these definitions, it is clear that a great deal of what goes on in the human mind is not memetic. First, perception and visual memory need not involve memes. You can look at a beautiful scene, or taste a delicious meal,

and remember them in detail without any memes being involved (unless you describe your experience using words, which are memes).

Second, not all learning involves memes. What you learn by yourself through classical conditioning (association) or by operant conditioning (trial and error) need not be memetic. Many other creatures are capable of these processes, and of extensive learning, but they do not have memes because they cannot pass on what they learn to anyone else. There may be a limited capacity for imitation in song birds, porpoises, dolphins, and possibly some primates (Heyes & Galef, 1996; Whiten & Ham, 1992). These species may, therefore, have some memes and some kind of culture (and there may be others we are simply unaware of). However, only humans are capable of widespread and general imitation, and only humans seem to find imitation rewarding in itself. As Meltzoff (1996) puts it, humans are the consummate "imitative generalist". It is this ability that makes a second replicator possible, and so leads to memetic evolution.

Third, not all communication is memetic. A baby's cry or a hearty laugh can be understood by anyone in any culture. In other species that are incapable of imitation, there can be complex communication, such as when bees indicate the location of food by dancing, or vervet monkeys alert others to one of several kinds of threat by using a different call. These examples entail true communication of information, but no copying of what is transmitted. There can be no evolution of the signals used except by a change in the genes. Only when a skill, behavior, story or other idea is copied from one person to another can we say that a meme has been replicated and a new kind of evolution can occur.

To understand this new kind of evolution we must apply the principles of Universal Darwinism, with the meme as replicator, the human being as a selection and copying system, and culture as the selective environment. It is important to remember that memetic evolution proceeds not in the interest of the genes, nor in the interest of the individual who carries the memes, but in the interest of the memes themselves. This is what it means to have a second replicator. This means that memetics differs fundamentally from sociobiology. Sociobiologists argue that the genes must always keep culture on a leash (Wilson, 1978). Similarly, evolutionary psychologists argue that the human mind evolved to solve the problems of a hunter-gatherer way of

life, and all our behaviors, beliefs, tendencies and customs are adaptations that ultimately come back to biological advantage (Barkow, Cosmides & Tooby, 1992; Pinker, 1997). But memetics suggests that once imitation evolved, and a new replicator was born, memetic evolution took off in directions dictated by memetic, not genetic, advantage. In the co-evolution of memes and genes, the two replicators would not always pull in the same direction (Blackmore, 1999).

From the meme's eye view the important question is why some memes survive and get copied into many brains or artefacts, while others do not. We might like to imagine that ideas that are good, useful, true, or beautiful, should survive in preference to those which are false, or useless. From the meme's point of view this is irrelevant. If a meme can survive and get replicated it will. Generally we humans do try to select true ideas over false ones, and good over bad; after all our biology has set us up to do just that, but we do it imperfectly, and we leave all kinds of opportunities for other memes to get copied—using us as their copying machinery.

Let's consider some examples of selfish memes that survive well in spite of being useless, false, or even harmful. At the simplest end of the continuum are self-replicating viral sentences such as "copy me" or "pass me on". As Hofstadter (1985) points out these are unlikely to be successful unless they are paired with an incentive or a threat (as in chain letters, for example). A threat such as "Say me or I'll put a curse on you!" is unlikely to be able to keep its word and few people over the age of five are likely to fall for it unless—Hofstadter adds—you simply tack on the phrase "in the afterlife".

This brings us to the other end of the spectrum—to vast groups of inter-related memes such as religions, political ideologies, scientific theories, and New Age dogmas. A group of memes that works together is called a 'co-adapted meme-complex' (Dawkins, 1976) or memeplex (Speel, 1995). Genes often group together in this way. For example genes for catching prey have coevolved with genes for digesting meat, while genes for grazing have coevolved with genes for digesting grass. These genes become linked and many animals have the combination. Similarly memes work together in the sense that a certain meme may do better in the presence of various other memes than it can on its own, and if the two can be linked the combination

may flourish. Such combinations can be self-replicating, self-protecting, and are usually passed on together.

Dawkins uses Catholicism as an example of a group of memes that have succeeded for centuries in spite of being false. For example, at Holy mass, the wine is supposed to turn *literally* into the blood of Christ, even though the wine still smells and tastes as it did before, and would not show up as Christ's blood in a DNA test. Millions of people believe in heaven and hell, an invisible and all-powerful God, the virgin birth and the Holy Trinity. Why? Because, says Dawkins, these memes have what it takes to get copied and to survive. The idea of God provides comfort, and an 'explanation' of our origins and purpose here on earth, and it cannot easily be overthrown by testing it. God is invisible, can see all your sins, and will punish you, but not until you are dead. Other memes directly discourage testing, such as the doctrine that faith is good and questioning is bad (the opposite of how it is in science). Other memes that protect the Catholic memeplex include exhortations to marry another Catholic and bring up lots of children in the faith, or to convert others. Giving money for the building and maintenance of great churches, cathedrals, and monuments inspires further meme hosts. In all these ways money and effort is diverted into the spreading of memes. The memes make us work for their propagation.

Memes such as religions, cults, fads and ineffective therapies, have been described as viruses of the mind because they infect people and demand their resources in spite of being false. Some authors have emphasized these pernicious kinds of meme and even implied that most memes are viral (for example Brodie, 1996; Lynch, 1996). However, memes can vary across a wide spectrum from being most like viruses, through being more like commensals or symbionts (living in peaceable harmony with their hosts), to being our most valuable tools for living (such as our languages, technology and scientific theories) (Blackmore, 1999; Cloak, 1975; Dennett, 1995). Without memes we could not speak, write, enjoy stories and songs, or do most of the things we associate with being human. Memes are the tools with which we think, and our minds are a mass of memes. As Dennett puts it, a person is "the radically new kind of entity created when a particular sort of animal is properly furnished by—or infested with—memes" (Dennett, 1995, p 341).

Note that successful memeplexes were not deliberately designed by anyone, but were created by the process of memetic selection. Presumably there have always been countless competing memes—whether religions, political theories, ways of curing cancer, clothes fashions, or musical styles—the point about memetic evolution is that the ones we see around us now are those that survived in the competition to be copied. They had what it takes to be a good replicator.

3. THE NATURE OF SELF

The most powerful and insidious of all the memeplexes is, I shall argue, our very own 'self'. It is perhaps hard to think of yourself as a memeplex, so I need to start by considering two major kinds of theory about the self.

On the one hand are what we might call "real self" theories. They treat the self as a persistent entity that lasts a lifetime, is separate from the brain and from the world around, has memories and beliefs, initiates actions, experiences the world, and makes decisions. On the other hand are what we might call "illusory self" theories. They deny the self any persisting coherence or efficacy, and liken it rather to a bundle of thoughts, sensations, or experiences tied together by a common history. On these theories, the illusion of continuity and separateness is provided by a story the brain tells, or a fantasy it weaves.

Everyday experience, ordinary speech and 'common sense' all seem to favor the "real self" but there are plenty of reasons to doubt it. First, as Hume (1739-40) observed long ago, introspection does not reveal a persistent self. Sit quietly and stare into your own experience and you find only experience—you do not also find a self who is experiencing. This kind of disciplined introspection is often used in meditation, especially in Buddhist traditions, to undermine the grasping (and ultimately the suffering) of the ordinary self. The Buddhist concept of no-self is not that there is no self at all, but that there is no persisting, unchanging self. As the Buddha explained, "actions do exist, and also their consequences, but the person that acts does not" (Parfit, 1984). Among modern philosophers, Parfit calls himself a 'bundle theorist' like the Buddha, and Strawson (1997) likens the self to a series of pearls on a string.

Modern neuroscience provides many other reasons for rejecting the 'real self'. First we may ask where it is. If you look inside a brain you do not see a central place where a self might live, and from where it might direct operations. You just see a lump of porridge-like stuff or, with magnification, millions of neurons connected in billions of ways to each other. Indeed the more you understand about what is going on in the brain, the less need there seems to be for a central experiencing self. Operations like perception and memory entail information moving rapidly in numerous parallel pathways. When you decide to pick up a cup of coffee there is no need for a self to oversee the action. Rather, there is processing in many areas of the brain at once, coordinating a detailed body image with incoming visual and tactile information, and ensuring a skilful lift.

As Dennett (1991) points out, there is no 'Cartesian Theatre' in the head, with a mental screen on which our images are projected; no central place to which all the inputs come in and from which the outputs go out. We may like to think of ourselves as a central perceiver and controller in charge but this is just a myth. He suggests instead that the brain constructs multiple drafts of what is going on, and just one of these is the verbal serial story we tell ourselves about the mythical self who is in charge. Claxton goes even further and concludes that consciousness is "a mechanism for constructing dubious stories whose purpose is to defend a superfluous and inaccurate sense of self." (1994, p 150).

What then is the role of consciousness in directing actions? None at all, I would suggest. Some experiments by Libet (1985) hint as much. He asked subjects to flex their wrists many times, whenever they felt like it—deliberately and spontaneously—and measured three things. First, the time at which they moved, second the time at which the 'readiness potential' in their brain showed that motor coordination was beginning and third, the time at which they decided to act. This latter he measured by asking them to state the position of a revolving spot on a clock face. If the conscious decision to act starts the process, then this should come first but he found it did not. The readiness potential preceded it by more than 300 milliseconds, a long time in terms of brain processing. Many different interpretations of Libet's findings have been given (see the many commentaries after the article), and Libet himself reserves a role for consciousness in vetoing actions once they have

begun. Nevertheless, the results should not be surprising. The idea that consciousness starts an action would be magic. We should not be surprised that conscious experience itself takes some time to build up.

Other experiments of Libet's (1981) show just this, that about half a second of continuous activity in sensory cortex is required before a person becomes conscious of a sensation. If this is so, then we must conclude that consciousness cannot be the receiver of impressions or the initiator of actions, as it feels as though it is. When we hear a sudden noise and jump, the jump begins before we are consciously aware of the noise. When we reach to catch a falling glass, we move before we are aware of the need. Consciousness does not do it. And our self does not do it.

These are just some of the reasons for rejecting the 'real self' view in favor of the idea that the self is a story about a powerful little person inside who does not really exist. To avoid misunderstanding I must make my view quite clear. There are human bodies and brains. Those brains can see, imagine, and think, and these processes may entail hierarchies, control mechanisms, and a body image. However, there is not, in addition, a central, persisting conscious self that receives the impressions or makes the decisions.

The implications for consciousness are this. The whole problem of consciousness stems from making a distinction between the world that is perceived and the self who is perceiving it, but if this self is just a myth, then this distinction must be false.

4. MEMEPLEXES AND THE SELFPLEX

If the self is a myth, then why are selves constructed at all? The answer, I suggest, is that the memes do it. They create a vast complex of memes, a self-memeplex, or selfplex. The selfplex may start out as a rather simple construct. We know that many other animals have a body image for coordinating actions, and some have some kind of sense of self. For example, most primates live in complex societies and are capable of deception, strategic planning, the formation of alliances and hierarchies, and Machiavellian intelligence (Whiten & Byrne, 1997). All this requires a clear sense of who is who and, presumably, a sense of oneself in the picture. Yet without language

the sense of self can get no further than this. Humans are capable of imitation and of language. So the scene is set for all kinds of language-based memes to jump in. The self now becomes a word to which can be attached desires, intentions, loves and hates, ambitions and fears. "I" love the Simpsons. "I" am going to be a famous artist. "I" believe in freedom of speech.

As we have seen, memes club together into memeplexes when the individual memes survive better in each other's company than they do on their own. This is true of the memes that make up the selfplex. Each of us comes across countless ideas every day but most are forgotten. However, any that become "my" belief are protected. I will fight for my beliefs; I will argue for them with others and so pass them on. The same is true of my plans for the future. Once I have got it into my head that I want to go to Bali for my holiday I will collect brochures, read books, and buy pictures of Bali. These memes spread better because they are part of "my" plans. The self becomes an idea to which are attached all sorts of verbal labels—nice, nasty, reliable, punctual, disobedient, friendly or sexy. Note that there is, of course, a body that behaves in certain ways and looks a certain way, but this is not how we talk about ourselves or each other. We speak about our 'self' as being the one who is nice or nasty. We don't just mean the body—we mean the inner 'me' who has this personality and is responsible for 'my' actions. As language and society become more and more complex we can say more and more things about this self, and it can desire more and more possessions and achievements.

In this way all kinds of memes succeed better because they become part of a self, and so the selfplex grows—and grows. This is how we come to acquire a story about a little self inside that *has* desires and plans, that *has* free will and the power to make decisions, and that *has* consciousness. Ordinary human consciousness is thus constructed by the memes, using the human meme machine. Our consciousness is the way it is because of the success of the memes that make up the selfplex.

5. ESCAPING ORDINARY CONSCIOUSNESS

If ordinary human consciousness is a construct of the memes, is it possible for a human being to drop all the selfish memes of the selfplex? If so,

what would it be like to be a meme machine without a selfplex? And why would people wish to do it?

First, it does appear to be possible. Many people work at letting go of the false self, for example through meditation, or by practices such as mindfulness or paying attention to the present moment all the time. There is no doubt that when the power of the self is undermined, or insight into its illusory nature is gained, the quality of consciousness changes. People describe it as becoming more open and spacious, or as though everything becomes more vivid and 'as it really is'. Apparently consciousness does not disappear, though in states of complete selflessness description is difficult and paradoxes abound.

Integrating the results of endeavours like this into psychology is not easy, yet some psychologists are trying to do so (for example Claxton, 1986; Crook & Fontana, 1990; Pickering, 1997). The memetic view may help. I have suggested that many of the contents of consciousness are not memes (such as immediate perceptions or emotions), while others are—including words and stories, and the false self with all its desires, possessions, and beliefs. Meditation and mindfulness aim at letting go of all the words and beliefs and opinions, while leaving the immediate sensations alone. In other words, they are practices that selectively weed out the memes. The awareness that remains is not based on memes and does not entail the false dichotomy between self and other.

Why should people want to do this? The answer depends on what you think the value of the selfplex is. For Dennett the memes that make up the self are indispensable tools for thinking with, and the user illusion is *benign*, but I believe it is not so benign. This false self is the centre of human suffering. It is the self who has endless wants and desires and is never satisfied, who is loved or admired or rejected, who gets rich or famous or disappointed. Perhaps it is possible, and even desirable, to dismantle the selfplex, and thus transform consciousness. I have described this as waking from the meme-dream (Blackmore, 1999).

If you take this view it is hard to see why anyone would want to construct an artificial system that had ordinary human-like consciousness, with all the suffering that entails. However, we can use this view to ask what would make such a construction possible.

6. ARTIFICIAL CONSCIOUSNESS

We may now return to the question I asked at the start. What would it take for an artificial system to have something like human consciousness? The answer I gave then should now make sense.

I have argued that ordinary human consciousness feels the way it does because the memes have constructed a vast selfplex around the idea of an inner perceiving self. This means that the only systems that can have consciousness like ours are systems that have memes. If you want to build an intelligent system with human-like consciousness, then you need to make it a meme machine. It has to be capable of imitation.

At present no artificial systems are capable of imitation in the way that humans are. Information is copied from one machine to another and we might ask whether this counts as memetic, or more generally whether it might sustain the evolutionary algorithm. For much information the answer is clearly no. Files are routinely copied from machine to machine but they are copied without errors. So there is no variation and therefore no possibility of evolution. However, there are many examples of evolution in artificial systems. The most obvious is the use of evolutionary algorithms.

Less obvious examples include the development of commercial computer software, and the evolution of ideas on the internet. Take, for example, a word processor like Word 6 with which this paper is being written. The code that makes it up has been copied millions of times and now exists in computers all over the world. Most copies are identical. However, Word 6 contains much code that was previously in Word 5 and in even earlier versions of Word. The selection pressures on this code came from the users who found Word useful, who created many documents using it, recommended it to their friends in preference to other word processors, and so on. In this example, the code acts something like a genotype, while the documents fulfil a role like that of the organism or phenotype (though we must cautious with such analogies, Blackmore, 1999).

The internet is a massive new arena for the evolution of ideas. Heredity occurs because messages can easily be stored and copied. Copying fidelity is high, but variation is introduced every time someone changes something in a message or combines it with something else. Selection is fierce because most

messages are ignored while some are actively sought out and passed on, and a few sit on web sites that are visited by millions of people. In this evolutionary system we may expect all kinds of memeplexes to evolve, as some memes thrive better in groups than they can on their own. Indeed web sites are a kind of memeplex, as are internet viruses and electronic advertisements. The hardware is also driven by this memetic evolution and in this sense the memes have forced us to create the internet for their own propagation. However, there is no scope for a selfplex to form because there is no organism with a body image, a rudimentary sense of self and the ability to imitate. At the moment there are only human users whose selfplexes just become more complex and overloaded as they use the internet.

These, then, are a few examples of evolution occurring in artificial systems. However, none of them involves anything like human imitation, so we should not expect them to give rise to a selfplex or anything like human consciousness.

Many attempts have been made to produce systems that behave more like human beings, and have human-like intelligence. The whole vast enterprise of artificial intelligence has been directed mainly at making artificial systems that can do things that, if done by a human, would be deemed intelligent. The focus has shifted greatly over the years as we have learned that some skills that first appeared very difficult (like playing chess for example) turn out to be relatively easy for artificial systems, while others that are easy for us (such as vision and natural language) turned out to be extremely hard to create artificially. However, no one has tried to recreate the human capacity for imitation. Even systems that are explicitly modeled on human abilities (like COG, for example) do not use imitation, even though human infants are capable of imitation almost from the moment they are born (Meltzoff, 1996).

I suggest that what makes humans unique is not our intelligence *per se*, not our ability to solve problems, or think logically, nor even our use of language, but our ability to imitate. Indeed I have argued that imitation came first and these other abilities followed (Blackmore, 1999). If this view is correct, then we will only create human-like intelligence, and human-like consciousness, when we give an artificial system a human-like ability to imitate.

What, then, does imitation entail? It is a surprisingly difficult process (even though it comes so naturally to us), and this may be why it has apparently appeared only patchily in the course of evolution so far, and only once in the general form found in humans. Imagine that I do some simple action and you copy me. Let's suppose I pick up my pen, flourish it in the air to spell out my name, and put it down again. I imagine that you would easily be able to copy that action with reasonable accuracy, but how? The process involved is a kind of reverse engineering. You must observe the action from your view point, construct some kind of model or copy of what has been done and store it, then convert that into a schema for action that you will perform and that will end up looking, to an external observer, like the same action. Among the difficult processes here are not only the memory and motor control, but the conversion of viewpoints, and the decisions about what has to be imitated.

This last is important. We would probably all agree on whether you did a good imitation of my action, but in doing so we have relied on all kinds of built-in systems for deciding what is important. For example, does it have to be the same pen? Will a pencil do? Does the exact angle of your arm matter, or the height of the writing in the air, or what you write, or even the spelling? Imitation is not a question of making exact copies of actions, but of making many decisions about what to copy. Looking at imitation this way we can see that much of our culture consists of ways of deciding what to copy and what not. Language is the prime example because words digitize the sounds we transmit and constrain what is passed on. Writing commits it to paper, and the variation in handwriting can be ignored when you know how to recognize a letter P or B. When we tell a story we have mutually agreed criteria concerning what counts as the "gist" that we would pass on if we repeated the tale to someone else. In this way much of human intelligence comes to look like ways of deciding what to copy and what not. We can now see that making an artificial system that could do this kind of imitation would be extremely difficult.

I am now going (conveniently) to ignore all these difficulties and imagine that it could be done. I shall consider two uses to which such artificial imitation might be put. First we might build robots to imitate humans. Second

we might make robots to copy each other. The consequences are rather different.

6.1 Robots Copying People

Imagine now a robot that can copy a human being (putting aside all the enormous technical difficulties of building it, and allowing it to have whatever perceptual and motor systems, sound producing systems or grammar modules that it needs to be able to do this). It has to be a robot, or virtual robot, because it must have a human-like body in order to imitate human actions. It now starts copying movements and sounds made by the humans around it. Soon it will pick up whatever language is being spoken around it, talking about the things the humans talk about, and beginning to use the word "I" in the way the humans do. It will pass on memes of all kinds as it imitates one human being and is observed by another. In time it will presumably start talking about itself just like the people do. As it does so, a selfplex will start to form, which includes the idea of an inner self who has beliefs and desires, who has free will and makes decisions, and who is conscious.

But it's just a robot, you might say. It must be a zombie without consciousness. But that would be to return to a magical idea of what consciousness is. Like Dennett (1991) I reject that view. Indeed I think our robot would become conscious in just the same way that we humans all became conscious during our lifetimes—by imitating others and so allowing the memes to construct a selfplex called "me".

If ever we were capable of creating such robots, should we do so? They would, I suggest, have all the joys and desperation, all the pleasures and sufferings, of ordinary human beings. The moral issues would be profound and we should certainly not undertake such a venture lightly. Perhaps happily, it is unlikely to be technically possible for a long time yet.

6.2 Robots Copying Robots

The other possibility is to create robots that can imitate each other. This is technically more feasible. The capabilities of the robots could be much more limited, so that the things they had to be able to copy could also be limited.

Nevertheless, imagining such robots—let's call them copybots—is an interesting thought experiment.

Let's imagine a group of them ambling about in some kind of relatively interesting and changing environment. Each copybot has a simple sensory system, a system for making variable sounds (perhaps dependent on its own position or some aspect of its sensory input), and a memory for the sounds it hears. Most importantly, it can imitate (though imperfectly) the sounds it hears. Or imagine the same principle but using gestures with a robot arm, movements of its entire body, or even visual displays on a screen. Now imagine that all the copybots start roaming around squeaking and bleeping, and copying each other's squeaks and bleeps.

The environment will soon become full of noise and the copybots will be unable to copy every sound they hear. Depending on how their perception and imitation systems work, they will inevitably ignore some sounds and imitate others. Everything is then in place for the evolutionary algorithm to run—there is heredity, variation, and selection—the sounds (or the stored instructions for making the sounds) are the replicator. What will happen now?

This is only a thought experiment but my guess is that the sounds will begin to evolve. Some sounds will be copied more accurately, some be more easily distinguished from each other, others more easily remembered (depending on various characteristics of the copybots). Patterns will then begin to appear. Some sounds would be made more often, depending on events in the environment and the positions of the copybots themselves. I think this could be called a language. If so it would not be the same as any language currently used by any natural or artificial systems. It would have evolved itself out of the copybots' ability to imitate.

If this worked, interesting questions would arise. Are the copybots really communicating? Are they talking *about* something? If so, symbolic reference would have arisen out of simply providing the robots with the capacity to imitate.

If this happened would we be able to understand them? Not unless we had perceptual and memory systems similar enough to theirs and could get in there with them and learn the way they did, by experience and imitation.

I have given the example of copying sounds because of its obvious implications for language evolution, but the same would apply to any actions

the copybots were capable of performing and copying. They might evolve strange dances, or forms of visual communication that we can scarcely imagine. If they had appropriate resources lying about, and built in needs (such as a need to acquire energy to keep going), then they might begin to copy ways of fulfilling those needs, or even develop some kind of technology. The evolution of human technology is possible only because we can copy what has been done before; similarly copybot technology might be possible precisely because of imitation.

These are wild speculations indeed but I raise them only to ask the question—would the copybots be conscious? My answer would be this. The copybots can imitate and so memetic evolution will get going and memeplexes form. The copybots acquire language (of sorts) through imitation and almost certainly this language would include terms of self-reference, because the most important things to talk about must be other copybots and, by implication, oneself. Around this core of self-reference a memeplex might form that has something in common with the human selfplex. Yet the differences would be enormous. We have evolved from ape-like ancestors, and have all the biological needs and complex social skills of apes. The machinery that does all our copying was created initially by the genes for their own benefit, and then adapted by the memes. The copybots are quite different in these important ways. So their memeplexes will be quite different too.

I conclude that the copybots would not have human-like consciousness, but would they be conscious at all? I suspect that if we could learn their language we might find them talking about their own perceptions, memories, thoughts and ideas, and so we might attribute consciousness to them just as we do to each other. And just as it is with each other, we could never really know what their consciousness was like.

If we were capable of creating copybots, should we do so? The situation is rather different from the human-copying robots, whose consciousness would necessarily be like that of a human to the extent to which they could really imitate us. The copybots, in contrast, would be copying each other and so letting loose a new evolutionary process, a new culture, and a new kind of consciousness. Although we might have created the copybots in the first place

we would neither be able to predict, nor control, the results of that evolutionary process.

REFERENCES

Barkow, J.H., Cosmides, L. and Tooby, J., editors. 1992. *The adapted mind: Evolutionary psychology and the generation of culture.* Oxford, England, Oxford University Press.

Blackmore, S.J. 1999. Waking from the meme dream, in: *The psychology of awakening: Buddhism, science and psychotherapy,* edited by Watson, G., Claxton, G. and Batchelor, S., Dorset, England, Prism Press.

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, England, Oxford University Press.

Brodie, R. 1996. *Virus of the mind: The new science of the meme.* Seattle, Washington, USA, Integral Press.

Calvin, W.H. 1996. *How brains think,* London, England, Weidenfeld & Nicolson.

Chalmers, D. 1996. *The conscious mind,* New York, USA, Oxford University Press.

Churchland, P.S. 1988. Reductionism and the neurobiological basis of consciousness, in: *Consciousness in contemporary science,* edited by Marcel, A.M. and Bisiach, E., Oxford, England, Oxford University Press, 273–304.

Claxton, G. 1994. *Noises from the darkroom.* London, England, Aquarian.

Claxton, G. Ed. 1986. *Beyond therapy: The impact of eastern religions on psychological theory and practice.* London, England, Wisdom; 1996. Dorset, England, Prism Press.

Cloak, F.T. 1975. Is a cultural ethology possible? *Human Ecology,* **3,** 161–182.

Crook, J. and Fontana, D. 1990. *Space in mind: East-West psychology and contemporary Buddhism,* London, England, Element.

Darwin, C. 1859. *On the origin of species by means of natural selection.* 1968. London, England, Murray and Penguin.

Dawkins, R. 1976. *The selfish gene,* Oxford, England, Oxford University Press. 1989. New edition with additional material.

Dawkins, R. 1993. Viruses of the mind. In *Dennett and his critics: Demystifying mind,* edited by Dahlbohm, B., Oxford, England, Blackwell.

Dennett, D. 1991. *Consciousness explained.* Boston, Massachusetts, USA, Little, Brown.

Dennett, D. 1995. *Darwin's dangerous idea,* London, England, Penguin.

Edelman, G.M. 1989. *Neural Darwinism: The theory of neuronal group selection,* Oxford, England, Oxford University Press.

Hameroff, S.R. and Penrose, R. 1996. Conscious events as orchestrated spacetime selections. *Journal of Consciousness Studies,* **3,** 36–53.

Heyes, C.M. and Galef, B.G., editors. 1996. *Social learning in animals: The roots of culture.* San Diego, California, USA, Academic Press.

Hofstadter, D. 1985. *Metamagical themas: Questing for the essence of mind and pattern.* New York, USA, Basic Books.

Hume, D. 1739-1740. *A treatise of human nature.* Oxford, England.

Libet, B. 1981. The experimental evidence of subjective referral of a sensory experience backwards in time. *Philosophy of Science,* **48,** 182–197.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences,* **8,** 529–539, with commentaries in: ibid., **8,** 539–566 ; ibid., **10,** 318–321.

Lynch, A. 1996. *Thought contagion: How belief spreads through society.* New York, USA, Basic Books.

Meltzoff, A.N. 1996. The human infant as imitative generalist: A 20-year progress report on infant imitation with implications for comparative psychology, in: *Social learning in animals: The roots of culture,* edited by Heyes, C.M. and Galef, B.G., San Diego, California, USA, Academic Press, 347–370.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review,* **83,** 435–450.

Parfit, D. 1984. *Reasons and persons.* Oxford, England, Oxford University Press.

Pickering, J., editor. 1997. *The authority of experience: Essays on Buddhism and psychology,* London, England, Curzon Press.

Pinker, S. 1997. *How the mind works,* London, England, Penguin.

Plotkin, H. 1993. *Darwin machines and the nature of knowledge.* Cambridge, Massachusetts, USA, Harvard University Press.

Speel, H.-C. 1995. *Memetics: On a conceptual framework for cultural evolution.* Paper presented at the symposium *Einstein meets Magritte*, Free University of Brussels, June 1995.

Strawson, G. 1997. The self. *Journal of Consciousness Studies*, **4**, 405–428.

Whiten, A. and Byrne, R.W. 1997. *Machiavellian intelligence II: Extensions and evaluations.* Cambridge, England, Cambridge University Press.

Whiten ,A. and Ham, R. 1992. On the nature and evolution of imitation in the animal kingdom: Reappraisal of a century of research, in: *Advances in the study of behavior*, Vol. 21, edited by Slater, P.J.B., Rosenblatt, J.S., Beer, C. and Milinski, M., San Diego, California, USA, Academic Press.

Wilson, E.O. 1978. *On human nature.* Cambridge, Massachusetts, USA, Harvard University Press.