CREATIVITY AND ARTIFICIAL INTELLIGENCE: A CONTRADICTION IN TERMS?

Margaret A. Boden

Abstract:

This is a philosophical question, not a scientific one. There are many examples of AI-systems whose products appear to be creative, and whose processing fits the criteria of combination, exploration, or transformation. (Arguably, that's strictly true of transformation only if the system is linked to contingencies external to the program.) But could any of them, or any conceivable computer-based system, be "really" creative?

That raises queries about whether "real" creativity must involve autonomy, intentionality, valuation, emotion, and consciousness. The answer, in each case, is yes (sometimes, with qualifications).

However, these concepts are hugely controversial in themselves, quite apart from their relation to creativity--and/or to AI. It follows that no clear answer can be given to the title-question. It remains open, until we have clear--and credible--accounts of all these matters.

I: Introduction

Many philosophers, and many otherwise hard-headed scientists too, deny the possibility of a computer's ever being creative. But that denial can mean two very different things.

Sometimes, such people are saying that a computer could not generate *apparently* creative performance. That's an empirical claim--and it's mistaken. There are many examples of seemingly creative artificial intelligence (AI) programs. Some of their novel results have been exhibited in major art galleries such as the Tate, the V&A, and MoMA; others have been awarded patents (in US law, only given for ideas not "obvious to a person skilled in the art").

It doesn't follow that AI-scientists will ever be able to engineer a neo-Chopin or neo-Mozart (although that goal has been approached more nearly than most people imagine: Cope 2001, 2006). It's even less likely that there will ever be an AI Shakespeare: the richness of the poet's world-knowledge and, even more important, the subtlety of his (and his readers') judgments of *relevance* will not, in practice, be matched (Sperber and Wilson 1986; Boden 2006: 7.iii.d). However, such individuals are the *creme de la creme.* Respectable, albeit lesser, examples of (apparent) AI creativity already abound (Boden 2004).

But sometimes, the sceptics are saying that *irrespective of its performance,* which might even match superlative human examples, no computer could "really" be creative. The creativity, we are told, lies entirely in the programmer. This is a philosophical claim, not an empirical one. On this view, it follows from the very nature of creativity that a creative computer is not a mere practical impossibility, but a contradiction in terms.

The case for that position can be made from opposite directions. On the one hand, it may be said that computers *possess* some some specific property which prevents them from being creative. On the other hand, it may be said that they *lack* some specific property/ies that humans

have, and which is/are necessarily involved in genuine creativity. The commonest candidate in the first class is being programmed (see Section III). Common candidates in the second class include autonomy, intentionality, consciousness, values, and emotion (see Section IV).

I have put the question in this way, i.e. speaking of "computers", because that is the way in which most philosophers (and others) express it. However, the question would be better put in terms of *computer-based systems, computing mechanisms,* or perhaps *information-processing systems* in general. That's because, for most people, the term "a computer" brings to mind a familiar, and relatively simple, type of machine--perhaps sitting on one's lap or the living-room table.

Granted, the "simplicity" of these familiar machines is only relative. The variety of information-processing functions carried out by today's computers is very much wider than most people believe. Not every current computing system is neatly GOFAI or connectionist (the only types of computation usually considered by philosophers). Even a desk-top PC harbours unsuspected complexities.

For example, some computational architectures cannot be simulated by a Turing machine (TM), because they have a number of interacting subsystems running concurrently and asynchronously. A humble PC has a number of concurrently active subsystems, controlling (and responding to) devices such as hard-drives, DVD-reader, internet connection, mouse, and keyboard. Each one delivers an "interrupt" signal when the relevant device is ready to process; whenever two or more interrupts happen to occur simultaneously, they may be handled according to pre-assigned priority rules, or they may be dealt with in random order. Less humble computer systems, such as those controlling an airliner, are even more complex. In general, computers that interact in some deep way with the many asynchronous changes happening in their physical environment will not behave like TMs, unless that environment is equivalent to a TM--which is unlikely to be true of chemical processes, weather systems, human brains, or the Internet. (These computers are very predictable most of the time, however, because computer scientists have developed ways of making them interact harmoniously and in accordance with specific rules.)

Nevertheless, and despite these often-ignored complexities, it may well be impossible--given current technology--to build *a single computer* with sufficient power to meet all the requirements for creativity that are outlined below. Perhaps we'd need thousands of them, linked in a tightly integrated network.

More to the point, our current computers, with their diverse hardware attachments (sensors, effectors, and communication channels), merely skim the surface of the huge variety of information-processing mechanisms--including chemical information-processing mechanisms--that are probably required for human-level, and much animal, intelligence. Future computer-based systems will doubtless be different in various ways. For instance, a pioneering computer model of the modulatory effects of diffusing chemicals in the brain reflects the fact that the *computational functions* of certain neurones may be temporarily altered, even though their *connectivities* remain unchanged (Smith et al. 2002). In other words, in a neuroscientifically realistic network, the anatomical pattern of neuronal connections *is not* all there is to it. Similarly (and *pace* critics of AI such as John Haugeland: 1978), the diffuse effects of chemically-induced moods, for instance, *are not* beyond simulation in computer models.

In short, the widespread sceptical intuitions about the creative limitations of computers *as we now know them* may well be largely correct. But they need not apply to all possible computer-based systems.

In Sections III to V, I address some of those sceptical intuitions. But first, we need some clarification of the notion of creativity.

II: What is creativity?

I said, above, that the denial of computer creativity can mean two different things. I should rather have said that it can mean *many* different things. For there is no universally accepted definition of creativity.

That's not surprising. Most concepts cannot be given hard-and-fast definitions that capture all the sub-varieties, and that make no arbitrary decisions about near-misses and borderline cases. To understand the concept, then, we need to focus on the similarities and differences between those varieties, near-misses, and borderline cases. In other words, dichotomous questions asking "Is this an X, *yes or no?",* and essentialist disputes about "What is X, *really?"* are usually misguided (see Sections IV and V).

However, if quibbling about spurious dichotomous distinctions is a waste of time, comparing alternative definitions is not. For it helps us to map the logical geography of the concept--or, better, the phenomenon--in question. The possible candidates for the attribution of *creativity* differ from each other in at least as many ways as do instances of *game,* or *chair* (Wittgenstein 1953). As a result, what people understand by the term--when they do try to give a hard-and-fast definition--can differ on a number of dimensions.

Some people, for instance, use it to mean the production of novel ideas and/or artefacts--*whether or not* these are valuable in any way, and (if so) *whether or not* their originator recognizes their value. On this view, a schizophrenic's word-salad is creative purely on account of its novelty. Its potential (for third-party use) as a source of poetic imagery is irrelevant--as is the person's own inability to recognize, and value, that potential. Likewise, Johannes Kepler is here seen as being no less creative when he declared his new idea about non-circular planetary orbits to be "a cartload of dung" than when, several years later, he realized its value ("Oh, what a foolish bird I have been!").

Others do include a value criterion (with or without a requirement that the originator recognize that value), so that mere novelty isn't enough to qualify for creativity. This implies that the identification--and explanation--of creativity is not a purely scientific matter. For science makes no value judgments. To be sure, it might be able to explain why we accept certain values: evolutionary psychology suggests why there is such a widespread tendency to prize shininess, for instance (Coss and Moore 1990; Coss 2003: 86-90). But showing that a feature is very useful for certain purposes (e.g. for life and/or evolutionary survival) is not, in itself, to show that those purposes are indeed valuable.

The value/s concerned in our judgments of creativity range very widely. Most are differentially

accepted by distinct social groups, so that disagreements on attributions of creativity arise accordingly. Some are relatively long-lasting, whereas others are subject to quick-changing fashions. Moreover, many are domain-specific: the values considered appropriate to judgments of creativity in physics, say, differ from those applied in painting, or poetry. Whether any values (symmetry, perhaps?) are universally accepted, and/or applicable to domains as different as physics and art, is a further contentious question (see Boden 2006: 8.iv.b).

Yet other people interpret "novelty" very strongly, so that if it turns out that Bloggs had the idea in question before Robinson did, then Robinson's idea wasn't truly creative after all. And as for Bloggs and Robinson themselves, some philosophers insist that no individual should be marked out by this honorific, because creative ideas arise within human groups (some of whose members are systematically overlooked in attributions of creativity, or even "discovery": Schaffer 1994).

Finally, if one takes the terms *ideas* and *artefacts* literally, then biological evolution cannot properly be called creative. Yet it's often regarded as a hugely creative process.

In short, the concept of creativity is highly slippery. Since it is used in so many very different ways, discussions often founder because the discussants are talking at cross-purposes. As remarked above, none of these can be pronounced "right" or "wrong": all reflect some aspects of the phenomena being discussed. Neverthless, it will be helpful here to pick one reasonably clear definition of the concept, with which to structure the argument.

My own definition sees creativity as the ability to generate creative ideas (a shorthand term, which includes artefacts)--where a creative idea is one that is *novel, surprising, and valuable.* We've just seen that "valuable" can be given multiple interpretations. But the other terms, also, have more than one meaning: "novel" has two, and "surprising" has three.

First, an idea may be new *to the originator;* in that case, it is psychologically creative (P-creative, for short). Or it may be, so far as is known, new *to the whole of human thought:* in that case, it is historically creative, or H-creative. Clearly, P-creativity includes H-creativity, because a historically new idea must also be new to the individual concerned. H-creativity is the more glamorous phenomenon, and (as noted above) is sometimes assumed to be essential to creativity properly so-called. For our purposes here, however, P-creativity is the more interesting. That's because it leads us to consider the psychological mechanisms that underlie originality--and which may, or may not, be simulated or even instantiated in computers.

Those mechanisms underlie the second aspect of my definition, namely, *surprise.* Phenomenologically, there are three types of surprise that we may feel on first encountering a creative idea (whether our own or someone else's).

One is the surprise we feel on seeing something happen which, because it is statistically unusual, we didn't expect--but which we always knew (or would have allowed, if we'd been asked) to have been possible. An example is the 100:1 outsider winning the horse-race. The second is the surprise of seeing something we didn't expect, and had never even considered--but which, once it arises, we can see to fit into some previously familiar pattern. For instance, a new painting or musical composition by an artist who has already mastered an established style. And

third, is the shock we experience when presented with a new idea that is seemingly not just improbable and/or unexpected, but downright *impossible.* (Closed-ring molecules, for instance, at the time when molecules were still thought of as being open strings.)

These three forms of surprise correspond to three types of creativity, generated by different psychological mechanisms (Boden 2004). The first is *combinational* creativity, which involves the unfamiliar juxtaposition of familiar ideas. (Think of poetic imagery, musical "quotations", or visual collage.) The generative processes, here, are those of associative memory.

The other two types of originality are closely related, since each springs out of a previously accepted style of thinking, or conceptual space. In *exploratory* creativity--which, on a first encounter, elicits the second type of surprise distinguished above--some familiar style/space is explored by generative processes that construct novel structures informed, and limited, by the constraints defining it. (Think of a novel fugue, or the synthesis of a new molecule within a known chemical family.) Sometimes, the exploration is deliberately aimed at discovering and/or exhibiting individual constraints, and even at testing their limits. Those sorts of exploration can be done not only by the creator, but by the 'audience' too. This is why the (structurally-based) *surprisingness* of a Bach fugue, for instance, may continue to engage us, even though the initial *feeling* of surprise is no longer aroused.

The third type is *transformational* creativity, wherein the novel structure does not fit into any known style--and even seems *impossible on first acquaintance. Typically, transformation follows on* conscientious exploration. Pure exploration gives way to transformation when the originator alters (or drops) some previously recognized constraint or adds one or more new ones. This enables *fundamentally* new structures to be generated, which were impossible relative to the previous (untransformed) style. The psychological mechanisms concerned are both exploratory (generative) and constraint-altering--where certain classes of alteration (e.g. constraint negation) can happen to many different constraints, in many different domains.

Transformation is not creation *ex nihilo.* Some of the previous generative constraints will remain; and the novel constraints may show traces of their pre-alteration ancestors. So even the most shocking novelties will be somehow related to the old style. This relationship may not be immediately apparent, still less well understood. Acceptance of the new structures as being "valuable" may take some time, for the stylistic similarities and differences must be both recognized and tolerated. (Pablo Picasso's cubist canvas *Les Demoiselles d'Avignon* was initially rejected even by his artist friends; when it was finally exhibited a few years later it caused a public scandal.)

In other words, if one includes value within the concept of creativity, then structural change in thinking-style is not sufficient for attributions of transformational creativity. Fresh judgments about value are required, since (by definition) some previously accepted/valued constraints will have been ignored. And some of the new value-judgments may depend on the extent to which the newly-transformed style can support further *exploratory* creativity of an interesting kind (see the comparison of two graphics programs in Section III).

III: What creativity-denying features do computers possess?

It's commonly said that computer creativity is a contradiction in terms because the computer is programmed. It does what it has been told to do--or, better, *what it has been empowered to do* --by its programmer. (The qualification is needed partly because the human being often cannot predict the machine's performance. Even more to the point, the program may enable the machine--perhaps in contact with the external world, including the Internet--to learn/develop/evolve preferences very different from, perhaps even antithetical to, those of the programmer.)

In other words, being programmed is the antithesis of being autonomous--which (so this objection runs) is a necessary feature of creativity. Whatever *genuine* creativity is involved in so-called computer creativity, then, must pertain only to the human programmer.

People making this objection may allow that combinational and exploratory creativity can at least be *simulated* by computers. But they typically draw the line at transformational creativity. For this involves not just a new thought but a new way of thinking--and that, it is claimed, cannot even be simulated by AI. After all, the sceptic will say, the rules/instructions specified in the program determine the computer's possible performance, and there's no going beyond them.

That remark is correct. But what it ignores is that the program may include rules *for changing itself.* For example, it may be able to learn--perhaps on the basis of unpredictable input from the environment, or perhaps due to its self-monitoring of internal 'experimentation' of various kinds. Or, more to the point for our purposes here, it may contain genetic algorithms, or GAs (see Boden 2006: 15.vi).

GAs can make random changes in the program's own task-oriented rules. These changes are similar to the point-mutations and crossovers that underlie biological evolution. Many evolutionary programs also include a *fitness* function, which selects the best members of each new generation of task-programs for use as 'parents' in the next round of random rule-changing. In the absence of an automated fitness function, the selection must be made by a human being.

Evolutionary programming can result in *prima facie* examples of transformational AI. For example, Karl Sims' (1991) graphics program produces images that often differ radically from their predecessors, with *no* visible family resemblance. That's possible because its GAs allow not only point-mutations (e.g. changing a numeral) within single programmed instructions, but also concatenations and/or hierarchical nestings of entire image-generating programs. So the program often arouses *impossibilist* surprise in the human beings observing it.

Whether Sims' computer system could deliver transformed *styles* as well as transformed *items* is another matter. For family resemblance is the essence of style. A style is a general pattern of ideas/artefacts that is sustained over time by the people adopting it. Sims' program cannot sustain a style, because a relatively fundamental mutation may occur at any time. Indeed, human selectors who try to steer the system towards certain colours or shapes are always frustrated: sooner or later, unwanted features appear. In brief, Sims' program is almost *too* transformational. Other evolutionary programs exist which allow only minor mutations, and correspondingly minor transformations. For instance, a GA program inspired by the sculptor William Latham can generate a sustained visual style (Todd and Latham 1992). However, this is new-ish rather than new, being highly reminiscent of the style of its human programmer/selector.

I said that various *prima facie* examples of transformational AI involve evolutionary programming. Why that cautious *"prima facie"?* Sims' program, after all, does generate radically transformed images. And Latham's program generates new-ish visual styles. Moreover, many highly efficient algorithms have been automatically evolved from *random* beginnings (for an early example, see Hillis 1992). If that's not transformation, what is?

Well, the objection here is a variant of the familiar argument that *A computer can do only what its program tells it to do.* It is posed by people who take the biological inspiration for evolutionary programming seriously (Pattee 1985; Cariani 1992; see also Boden 2006: 15.vi.c). They point out that programs are abstract systems, and as such are logically self-contained. Even evolutionary programs, like those discussed above, are essentially limited to the possibilities inherent in the GAs and other rules supplied by the programmer. Genuine, truly radical, transformations can arise in an evolving system, these objectors argue, only if it interacts *physically* with actual processes in the outside world.

Their favourite example concerns the origin of new organs of perception. They allow that once a light-sensor has arisen in a biological organism, it can become more powerful as a result of genetic mutations that can be approximated in AI programs. So an inefficient computer-vision system might, thanks to GAs, evolve into a better one. But the *first* light-sensor, they insist, can arise only if some mutation occurs which causes a bodily change that happens to make the organism sensitive to light for the very first time. The light--considered as a physical process--was always out there in the world, of course. But only now is it 'present' for the organism. One might say that only now has it passed from the world into the environment. That acceptance of light as part of the organism's environment depends crucially on physical processes--both in the world and in the living body. And these processes (so the argument goes) have no place in AI.

These sceptics (and many others) assume that AI is either simulation in a purely virtual world, or abstract programming that defines all the interactions that can happen between program and world--as in computer vision, for example. It follows, they say, that the generative potential of a computer program is constrained accordingly. So although much-improved artefacts can result from evolutionary computing, no fundamentally new capacities can possibly arise. For instance, if the physical parameters foreseen by the programmer as potentially relevant don't happen to include light, then no artificial eye can ever emerge. In general, there can be no *genuine* transformations in AI.

That may be true of AI systems that are purely virtual simulations with no representation of physical parameters--as both Sims' and Latham's programs are. But (for reasons explained below) it need not be true of all virtual simulations. And it's demonstrably not true of all AI systems--in particular, of some work in so-called "embodied" AI. Indeed, recent research in this area has resulted in the evolution of *a novel sensor:* the very thing which these critics claim can happen only in biology.

In brief, a team of researchers were using a GA to evolve oscillator circuits--in hardware, not in simulation (Bird and Layzell 2002). To their amazement, they ended up with a primitive radio receiver. That is, the final (automatically selected) circuit acted as a radio antenna--a "radio wave sensor"--that picked up and modified the background signal emanating from a nearby PC

monitor.

On investigation *post hoc*, it turned out that the evolution of the radio-wave sensor had been driven by unforeseen physical parameters. One of these was the aerial-like properties of all printed circuit boards, which the team hadn't previously considered. But other key parameters were not merely unforeseen but unforeseeable. For the oscillatory behavior of the evolved circuit depended largely on accidental--and seemingly irrelevant--factors. These included: spatial proximity to a PC monitor; the order in which the analogue switches had been set; and the fact that the soldering-iron left on a nearby workbench happened to be plugged in at the mains.

If the researchers had been aiming to evolve a radio-receiver, they would never have considered switch-order or soldering-irons. Nor would either of these matters necessarily be relevant outside the specific (physical) situation in which this research was done. On another occasion, perhaps, arcane physical properties of the paint on the surrounding wallpaper might have played a role. So we can't be sure that even research in *embodied* AI could confidently *aim* to evolve a new sensor. The contingencies involved may be too great, and too various. If so, doubt about (non-accidental) genuine transformations in AI still stands. But that they can sometimes happen unexpectedly is clear.

As suggested above, unexpected transformations might even happen in a *purely virtual* simulation, provided that it modelled the relevant interactions with the physical environment. For example, when Sims (1994) used GAs to evolve the anatomy and behaviour of creatures made up of 'blocks' of varying shapes and sizes, he found that some moved in lifelike ways whereas others, although equally efficient, did not. Their style (sic) of movement was hugely different from anything seen in biology--indeed, it was biologically *impossible.* On inspection, it turned out that Sims had mistakenly omitted a specific physical parameter from the simulated physics, and some creatures had evolved to take advantage of this. When he added it, and restarted the experiment, no non-lifelike behaviour appeared. So environmentally-based transformations might be in principle impossible in a purely virtual model *only if* they depend on specific features of the physical world that could never be included, because they resist algorithmic simulation.

In sum, it's true that *insofar as* a computer's performance is caused by its program, everything it does was somehow implicit in the instructions (perhaps including GAs) provided by its programmer. Only insofar as it can be affected by unforeseen events can genuinely new types of result emerge. Many of those events will be external to the system itself. However, they might be cultural/semantic as well as physical: accidental interactions with text or imagery on the Internet, for instance, might occasion creative transformations very different from what the programmer envisaged (see Boden forthcoming.) Moreover, they could also include internal events, if the system could perform information-processing internally, playing around in various ways (e.g. making novel combinations and exploring existing styles) with the structures already present in it.

IV: What essential features of creativity might computers lack?

We've already seen that the concept of creativity is complex, and highly slippery. But the kinds of slipperiness identified in Section II are only part of the philosophical story here. Creativity is

normally taken to be a property of new ideas and artefacts, and/or of the people who originate them. Now, we must ask whether it involves features which computers--or, better, computer-based systems (see Section I)--lack.

Specifically, does it imply autonomy, intentionality, consciousness, value, and emotion--all of which are commonly assumed to be denied to computers?

Two different questions arise here. One (the topic of this Section) asks whether each of those concepts is indeed crucial to creativity. The other (addressed in Section V) asks whether each is indeed necessarily lacking in computers.

A caveat: These questions are posed in essentialist terms *("crucial to* creativity", and *"necessarily* lacking"). Yet in Section I, I repudiated essentialist approaches. Hard and fast definitions, generating dichotomous distinctions, aren't available for any concept of interest, and certainly not for creativity--nor, in turn, for autonomy, intentionality, consciousness, value, and emotion. It follows that each of the questions expressed above can be misleading. Nevertheless, they will be helpful in structuring our discussion--much as focussing on a particular definition of creativity can be helpful too.

The notion of *autonomy,* or self-direction, is implicit in talk of someone's "originating" an idea. Indeed, creativity is often thought of as a species of freedom. This need not be understood (although it often is) as freedom in a strongly individual sense. We saw in Section II that some philosophers stress the social sources of originality, and are wary of ascribing the creative responsibility to individuals. (For a psychologist's defence of this position, see Csikszentmihalyi 1999.) Even so, each member of the group of interacting, and perhaps deliberately cooperating, people is assumed to share the general human property of freedom, or autonomy.

My own definition of creativity (in Section II) tacitly assumed that the novel idea was freely generated by the person concerned. Or rather, it assumed this with respect to H-creative ideas. The case of P-creative ideas is sometimes rather different. Sometimes, the origin of a P-creative idea can be predicted by a third party, or even largely brought about by their influence. Think of Socratic dialogue, for instance. Here, an idea that has never occurred to the pupil is deliberately eased into existence by the tutor. There's a real sense in which the student *was not* wholly autonomous in coming up with the P-creative idea.

In the normal case, however, the person realises the (perhaps very old and/or very widespread) idea for the first time without such direct intervention. The P-creative idea may, of course, be suggested/triggered by some environmental cue. But that's just to say that the thought isn't random--not that it isn't autonomous.

In short, all H-creativity, and almost all P-creativity, is autonomous. It's made possible partly by idiosyncratic personal experience, and partly by powerful (including meta-cognitive) information-processing mechanisms provided by evolution. The same is true of "free choice" in practical and moral reasoning (see below). In short, and despite all the many sociocultural (and physical) influences on thinking, creative ideas are freely generated.

What of intentionality (alias aboutness, or meaning)? This is, by definition, an essential

property of *ideas.* And it's essential to *artefacts,* too. For an artefact is not merely a physical thing many of whose features have been caused by a human being: that description would fit a collapsed tomato, accidentally trodden underfoot. Rather, it is a physical thing whose existence and/or some of whose features have been brought about by human action--typically, by deliberate human action. (Drop the "human", here, and one could allow that artefacts might be made by the actions of chimps, or by tool-using crows: Weir et al. 2002; Kacelnik et al. 2006.)

Some artefacts, to be sure, are not brought about by deliberate action: many poems and paintings of the Surrealists of the 1920s, for example. These artists engaged in automatic writing and painted while in trance states, in order to prioritize the unconscious mind--which Andre Breton declared to be "by far the most important part [of our mental world]". Indeed, Surrealism was defined by Breton as: "Pure psychic automatism by which one proposes to express ... the actual functioning of thought, in the absence of any control exerted by reason, exempt from all aesthetic or moral preoccupations" (Breton 1969). Even if one accepts this description at face value, however, one must grant that these artefacts arise from intentional, albeit involuntary, processes--namely, unconscious ideas and/or "automatic" (but meaningful) psychological processes.

So if creativity is defined--as it usually is--by reference to ideas and artefacts, it must be intentional. (Biological evolution could not then be classed as truly creative, despite the many near-impossibilist surprises in phylogenesis: see Section II, above. For it produces new organs and new species, not new artefacts; and it does this by means of non-intentional processes.)

The third item in our list of key concepts here is consciousness. What of that? Does creativity necessarily involve consciousness?

Well, one might doubt this. For it's not only the Surrealists, with their explicit stress on the artist's unconscious ideas and automatisms, who have suggested that consciousness *is not* essential to creativity. It has long been noted that creative ideas are often generated without consciousness--much to the surprise of the person concerned.

For instance, they may 'pop up' during artistic work or problem-solving without the originator's being able to say how this happened. Or they may arise in the so-called incubation phase of creativity, while the person is consciously thinking about something entirely different, or even asleep (Boden 2004: 29-35, 256-261). Moreover, they are *always* generated without full consciousness of the processes involved. But so too are the sentences we use in everyday conversation, and the perceptual, motor, and intellectual judgments that we make. (That's hardly surprising: if all such data, on every processing-level, were consciously available, they would present a paralysing information overload.) The partial unconsciousness of creativity isn't a special case--although people often speak of it, with awe, as though it were.

Why, then, is it so widely believed that creativity requires consciousness? The reason, I suggest, is that many of the creative activities of adults (in arts, crafts, and science) are consciously monitored throughout, sometimes in a highly self-conscious way (Harrison 1978). Even more to the point, they typically involve a conscious judgment that the final idea is valuable. Indeed, that's why some people are willing to ascribe "creativity" only if the creator recognizes the value of their own idea--as Kepler, at first, did not (see Section II); and as young

children, despite their high P-creativity, in general do not. When Pablo Picasso declared proudly *"Je ne cherche pas, je trouve!",* he was denying only sustained deliberation (and error) on his part, not conscious appreciation of the *trouvaille.*

This consideration suggests that the fourth item on our list is included within the third. In other words, valuation is one of the many species of consciousness. According to the Surrealists (and their inspiration, Sigmund Freud), valuation is sometimes unconscious. But even they allowed that their unconsciously generated artworks attracted their conscious approval once they had been formed. Breton (1937: 31ff.), for instance, spoke of the artist as a vigilant watchman *(guetteur)* in preconscious waters *(les eaux pre-conscientes).* Whether or not someone explicitly includes positive value in their definition (as I do: see Section II), people discussing creativity are in practice interested only in those novel ideas which someone approves. In brief, positive value is normally treated as an essential aspect of the concept.

Similar remarks apply to *emotion,* the last item on the list. I don't know of anyone who explicitly defines creativity with reference to emotion. My own intuition, here, is that a P-creative idea might perhaps be produced, and might even be recognized as valuable by its originator, without any specific emotion being involved. If that's so, then identifiable emotions (as distinct from the positive affect involved in judging something to be "valuable") aren't strictly necessary for creativity.

However, creativity in the arts is often associated with deep, and deeply personal, emotions. A poet's, or an architect's, grief at losing a lover may inspire a masterpiece that's admired for hundreds of years. And emotions of many different kinds appear to be involved in other cases of artistic creativity. Nor is scientific creativity immune to emotion. Perhaps it's rarely driven by grief, but it's sometimes driven by jealousy, even if of a relatively superficial kind. And it's often inspired by a passionate desire for recognition or monetary gain. The selfless quest for knowledge is less common than scientists like to make out. Perhaps intellectual curiosity drove Archimedes, and led him to leap from his bath in excitement on originating his novel, and highly valuable, idea--but even he expected to be rewarded by the king.

Indeed, positive valuation normally involves some emotional tone, howsoever mild that may be. And many creative thinkers have reported strong emotions as an aspect, even a driving aspect, of their work. So creativity typically, if not necessarily always, involves emotion.

As that word "driving" suggests, an emotion *is not* purely a matter of conscious feeling. To the contrary (as we'll see in Section V), emotions involve complex mechanisms deeply embedded in our mental architecture, and are needed for scheduling activities in complex multi-motivated creatures (Boden 2006: 7.i.d-f). Grief, for example, and its gradual assuagement through mourning, involves obsessional memories, sudden interrupts, and frequent distractions of goal-driven activities. However, it also involves intense feelings of sadness, desolation, and loss. It is these conscious aspects of emotion, or emotional qualia, which people normally have in mind when they speak of emotions in creative thinking. In the context of our discussion, then, emotion--like value--is a species of consciousness.

In sum: autonomy, intentionality, and consciousness (including valuation and emotion) are indeed--with minor qualifications--typical of creativity. It's not surprising that philosophers of an

essentialist disposition regard them as *essential* to creativity.

V: So could a computer really be creative?

To ask whether a computer could "really" be creative is to ask whether any of the concepts considered in Section IV is forever barred to computers (or to computer-based information-processing systems). The answer is clear in only one case: that of autonomy.

The presumption of autonomy as a *sine qua non* of creativity is what drives the objection that a computer can do only what its programmer tells/enables it to do. That objection was broadly accepted above (in Section III)-- *although* it was also pointed out that a computer's performance may be unpredictable, and may even involve the development of values/preferences very different from those of the programmer. So we must allow that, *in that strictly limited sense*, no programmed system can be truly autonomous. In that sense also, then, computers cannot be truly creative either.

This claim might be challenged on the grounds that some AI scientists--and some computer artists, too--actually make a point of describing their systems as "autonomous". In saying this, they are highlighting certain interesting features of the ways in which their machines function. That is, they are all (reasonably) distancing their computer's performance from an unalterable program written by a human programmer, and thus (reasonably) asserting some degree of independence on the machine's part. But they are not all focusing on the same features, so are using the term "autonomous" in three very different senses, to denote distinct types of processing--only one of which is at all analogous to human freedom (Boden 2010).

Fortunately, we need not enquire, here, just how close that analogy is--"fortunately", because the concept of freedom is itself highly contentious. Like most cognitive scientists, I interpret it in computational terms, as marking a particular type of cognitive/motivational complexity (Boden 2006: 7.i.g; Dennett 1984). Some philosophers would regard this view as not merely mistaken, but profoundly absurd (see below). Yet if we allow, as argued in Section III, that programmed systems do not possess autonomy--because the program, *even if* it's one which bears some resemblance to human freedom, was originated by a human being--that philosophical disagreement can be ignored.

Comparable philosophical disagreements arise with respect to the other key concepts here, however. And they cannot be sidelined so quickly. Intentionality and consciousness (and values and emotion, too) are hugely controversial. Indeed, they are so problematic that we don't understand them well enough to be able to ask the question about "real" creativity sensibly, never mind answer it.

Some philosophers will disagree. They deem it obvious, even near-axiomatic, that computers lack these key properties. This is true of writers in the broadly neo-Kantian tradition, from phenomenologists to postmodernists--recently, including some analytical philosophers too (e.g. McDowell 1994; Morris 1991; 1992). Martin Heidegger, for example, glossed intentionality and consciousness as *Dasein,* which he ascribed only to human beings. For orthodox Heideggerians, even non-human animals lack *Dasein* (a point disputed by some of his more scientifically-

minded followers--Wheeler 2005: 157-160). The notion that it could be possessed by computers is rejected by neo-Kantians as *a fortiori* absurd. So is the notion that intentionality could ever be explained by science--even by evolutionary theory and/or neuroscience. They argue that intentionality is the ground of all our conceptual thought, science included, so a naturalistic psychology is impossible (Boden 2006: 16.vi-viii).

Notoriously, the split between neo-Kantian and naturalistic/empiricist views is the deepest split within Western philosophy. That's true *even though* analytically trained philosophers now take neo-Kantianism more seriously than they did twenty years ago (cf. Williamson 2007). Not only does it affect discussion on all philosophical questions (cf. Blackburn 2005), but it cannot be definitively settled: there is no knock-down argument on either side.

Jerry Fodor, responding to John McDowell's version of the anti-naturalist position, declared that McDowell was "as good a contemporary representative of this philosophical sensibility as you could hope to find", but insisted that: "it's all wrong-headed. Science isn't an enemy, it's just us" (Fodor 1995: 8). Praising McDowell's book for raising "a number of our deepest perplexities", he defiantly added "Which, however, is not to say that I believe a word of it" (Fodor 1995b: 3). Notice that telltale expression, *I [don't] believe a word of it:* Fodor was admitting that he couldn't actually disprove McDowell's account. Certainly, he had identified several aspects of McDowell's core concept of "second nature" that he felt were mistaken, unjustified, or merely metaphorical. (What is it, for instance, to "resonate" to meaning?) And he'd offered his own, alternative, claims. But even the supremely self-confident Fodor didn't suggest that these were strictly provable.

One brave--or perhaps foolhardy?--philosophically-minded computer scientist, Brian Cantwell Smith (1996), has tried to bridge the gap by rolling intentionality and computation together within his basic metaphysical definition. The philosopher Haugeland is quoted on his dustjacket as saying: "Smith recreates our understanding of objects essentially from scratch--and changes, I think, everything." If Smith is right, then that is true. But it's by no means clear that he is right. (My own view is that Smith helped himself to the "dynamic flux", his version of Kant's noumenal world, without proper licence--see Boden 2006: 16.ix.e. He claims that he's pulled this concept up by its own bootstraps, in his final 60 pages, to form a "constructivist" metaphysics of objects and intentionality--cf. also pp. 188f. But it seems to me that he begs this fundamental philosophical question instead of answering it.)

Nor is this naturalist/neo-Kantian split the only problem here. On each side of the split, people disagree among themselves. The naturalists, for instance--whom I, like Fodor, believe (sic) must be basically correct--offer several radically different accounts of intentionality.

A few, like Smith, offer highly eccentric theories wherein intentionality and/or information is made metaphysically basic--i.e. not confined to minds, whether human or animal (e.g. Chalmers 1996). But even the less eccentric approaches vary significantly. For example, some naturalists define intentionality as a causal phenomenon (Dretske 1984, 1995), others gloss it as computational (Sloman 1986, 1987b), while yet others see it as somehow rooted in biological evolution (Dennett 1969; Millikan 1984; Papineau 1987).

None of these positions is free of difficulties. The evolutionist Ruth Millikan, for instance,

argues that a miraculously assembled molecule-for-molecule replica of a person (the so-called "swamp-man"), with causal powers identical to those of a real human being, would lack intentionality. Although the swamp-man's verbal responses would be just like ours, his/its words would lack all meaning--simply because he/it had been nano-assembled, not biologically evolved. My own view (as someone sympathetic to Millikan's approach) is that this highly counter-intuitive claim cannot be contradicted but may be ignored, much as in practice we ignore the theoretical possibility--according to statistical thermodynamics--of there being, if only for a split second, a snowball in Hell. Nevertheless, it must be admitted that swamp-man is something of an embarrassment. And swamp-man isn't the only problem. Causal theories of intentionality, for instance, have difficulty accounting for non-veridical content. In short, there is no theory of intentionality which satisfies all naturalists.

But at least those disputatious philosophers can understand each other. The situation is even worse with respect to consciousness. Not only do the naturalists and neo-Kantians locate it on opposite sides of the split, asking very different questions accordingly, but those gathered within the naturalist camp are far from mutual intelligibility, never mind agreement (Boden 2006: 14.x-xi; 16.iv.b). This is especially true with regard to what David Chalmers (1996) has called the "hard" problem, namely, the analysis/explanation of conscious sensations, or qualia.

The competing naturalist accounts of consciousness include a number that are expressed in relatively traditional terms (e.g. Block 1995, 2001; Searle 1992), and even more that appeal to science. Some recent analyses lean heavily on neuroscience (e.g. Baars 1988, 2001; Dehaene and Naccache 2001; Edelman and Tononi 2000; Edelman and Seth 2009), while others are based in speculations about quantum physics--thus inviting the charge of attempting to solve one mystery by citing another (e.g. Chalmers 1996; Penrose 1989, 1994; Walker 2000). One philosopher of strong scientific sympathies argues--a position that is possibly correct, but in my view unnecessarily defeatist--that the mystery here is permanent, because the human mind simply lacks the cognitive capacity to achieve a scientific understanding of consciousness, much as dogs lack whatever's needed to understand language or physics (McGinn 1991).

Besides all those candidates is a clutch of scientificially-influenced theories which strike many people, including other naturalists, as even more counter-intuitive than swamp-man. They're regarded as especially bizarre, even as unintelligible, when they're applied to qualia. These are the various computational theories of consciousness. Not all of them focus on the key question of this Section, namely, whether there can be such a thing as "machine consciousness" (Holland 2003). However, anyone who offers a computational account of (any aspect of) human consciousness must admit that suitably similar computer-based information-processing systems could, in principle, be conscious too.

Two of these theories are especially interesting: the analyses of human and animal consciousness developed over many years by Marvin Minsky (1985, 2006) and by Aaron Sloman (1999, 2000, 2010a,b; Sloman and Chrisley 2003). Both focus on the computational architecture of the mind as a whole, and they are similar in a number of ways. But Sloman's account is more systematic than Minsky's (and is closely related to his discussions of philosophical problems such as cause, freedom, possibility, reference, and intentionality).

Sloman points out--what is often stated, but also often forgotten--that the noun

"consciousness" is highly misleading. We'd do far better to consider cases where a subject is "conscious of" X or Y. There are many such cases, in both humans and animals, which differ considerably from each other. They don't differ merely in intensity, and nor do they fall on a continuous spectrum. Rather, they differ in their essential computational structure. In other words, the multitudinous varieties of consciousness are aspects of the multi-dimensional virtual machines which we call minds.

Like virtual machines in computers (Sloman argues), these aspects of mind are real, and have real causal effects. Qualia, for example, are internal computational states that have various effects on behaviour and/or on other aspects of the mind's information processing. They can exist only in virtual machines of significant structural complexity (he outlines the types of computational resources that are required). They can be accessed only by some other parts of the particular virtual machine concerned, and do not necessarily have any behavioural expression. In particular, they cannot always be described (by higher, self-monitoring, levels of the mind) in verbal terms. So whereas some computationalists--such as Daniel Dennett (1991, 1995)--deny the reality of qualia *even in human minds,* Sloman does not. (This doesn't mean that he identifies them with brain processes. Some computational states cannot be defined in the language of physical descriptions--even though they can exist, and have causal effects, only when implemented in some underlying physical mechanism.)

We saw in Section IV that the aspects of consciousness which are closely related with, and perhaps even essential for, creativity are emotions, value-judgments, and deliberate self-monitoring. According to Sloman, in order to generate such phenomena the virtual machine which is the mind needs to be of a certain kind.

Emotions, for instance, involve scheduling mechanisms that are necessary in multi-motive creatures acting in a complex and largely unpredictable world (Sloman 1978, 1987a, 1993). (So the emotionless Mr. Spock of *Star Trek* is an evolutionary impossibility.) That's true even of grief, which has driven every Renaissanace *Pieta* and many other works of art besides. This emotion, with its characteristic and diverse effects, can--and inevitably will--arise only in minds architecturally capable of deep personal love (Fisher 1990). Sloman has explained the psychology of this phenomenon, and its alleviation through mourning, in computational terms (Wright et al. 1986).

More accurately, grief and other emotions have been broadly outlined by him (and by Minsky: 2006) in computational terms, using theoretical concepts and insights drawn from AI. Functioning computer models based on his and Minsky's ideas are very few on the ground. Indeed, most so-called computer models of emotion are based on very superficial psychological theories. These normally assume, for instance, that emotion must have behavioural effects (which precludes allowing that grief can endure for years even though it is often temporarily dormant, or seemingly eclipsed by mirth). However, a few varieties of the emotion of anxiety have been simulated by Sloman's group (Wright 1997; Wright and Sloman 1997; see also Boden 2006: 7.i.f).

The types of anxiety concerned are among those which a nursemaid might experience while left in sole charge of a dozen hungry, attention-demanding, and active babies--in a nursery whose two open doors lead onto a busy road and a garden stream. Crucially, "might experience" here

doesn’t primarily mean "might feel". (In other words, this is not intended as a model of qualia.) For these distinct species of anxiety enable the nursemaid to prioritize her motives and schedule her actions appropriately.

They reflect the facts that some goals cannot be pursued simultaneously (she has only two hands, after all, and can’t be in two places at once); some goals conflict in a stronger sense, so that at most one of them can be achieved; some are hugely important, but never urgent; others are both important and urgent; some are relatively unimportant, but must be attempted urgently if they are to be attempted at all; some goals may be postponed (perhaps only for a limited period) while another is given priority; and others, which cannot be postponed indefinitely, must be abandoned if they can’t be achieved quite soon.

Clearly, these facts apply to all tasks that require multi-motive scheduling, where conflicts between motives can arise for many different reasons. So this isn’t merely a model of (some types of) anxiety. It illustrates the nature of emotion in general.

That’s not to say that it fully reflects the architectural complexity of the human case. The emotional demands on the simulated nursemaid, who has only seven motives to follow, are much more complex than those represented in other current simulations of emotion. To that extent, they are a significant advance. But real nursemaids, besides satisfying the seven motives represented in this model, have to worry also about cuddling the babies, bathing them, changing them, singing to them, protecting them from live electric plugs ... and so on.

The psychology of grief, and of many other emotions, is more challenging still. The necessary structural complexity is orders of magnitude greater than that of any current computer system. Even more to the point, it involves architectural distinctions which we are only just beginning to understand. So to attempt even to model (never mind instantiate) grief in a computer-based system today would be hugely premature.

It follows that a computational understanding of how some examples of human creativity are driven by grief is available to us only in the sketchiest terms. Nevertheless, these ideas help us to understand how grief is possible--and how among its many manifestations may be a drive to commemorate the lost loved one in painting, poetry, or song.

VI: Conclusion

The previous few paragraphs will strike some readers as ridiculous. For the notion that there could be a *computational* theory of consciousness (including emotions and valuation) seems, to many people, to be intuitively absurd or even unintelligible. (It’s worth mentioning, however, that Sloman’s account of grief appeared in a journal whose editor, as a consultant psychiatrist, is all too familiar with the ravages of grief and mourning.) As already remarked, that reaction can occur even on the naturalist side of the philosophical fence--and is *de rigueur* on the neo-Kantian side.

To reject computationalism, however, is not to agree on an alternative. As we saw in Section V, even the non-computationalist naturalists differ hugely about the nature of intentionality and

consciousness. Should one appeal to neuroscience, or to quantum physics? Or to neither? Should one favour a maverick philosophical position in which intentionality/information is metaphysically basic? Or should one throw up one's hands in despair, proclaiming that these matters lie forever beyond human ken?

In sum, the question whether a computer could ever "really" be creative is currently unanswerable, because it involves several highly contentious philosophical questions. If we take the argument about autonomy seriously, then we can agree that "AI-creativity" is a contradiction in terms *even though* a computer's performance may be very much more independent of its program than is usually assumed. But if we appeal, rather, to intentionality or consciousness, the question must remain open.

References

Baars, B. J. (1988), *A Cognitive Theory of Consciousness* (Cambridge: Cambridge University Press).

Baars, B. J. (2001), *In the Theatre of Consciousness: The Workspace of the Mind* (Oxford: Oxford University Press).

Bird, J., and Layzell, P. (2002), 'The Evolved Radio and its Implications for Modelling the Evolution of Novel Sensors', *Proceedings of Congress on Evolutionary Computation*, CEC-2002, 1836-1841.

Blackburn, S. W. (2005), *Truth: A Guide for the Perplexed.* The Gifford Lectures 2004 (London: Allen Lane).

Block, N. (1995), 'On a Confusion About a Function of Consciousness, *Behavioral and Brain Sciences*, 18(2): 227-287.

Block, N. (2001), 'Paradox and Cross Purposes in Recent Work on Consciousness', *Cognition*, 79: 197-219.

Boden, M. A. (2004), *The Creative Mind: Myths and Mechanisms.* 2nd edn., expanded/revised (London: Routledge). First edn. London: Weidenfeld & Nicolson, 1990.

Boden, M. A. (2006), *Mind as Machine: A History of Cognitive Science* (Oxford: Clarendon Press).

Boden, M. A. (2010), 'Autonomy, Integrity, and Computer Art', in M. A. Boden, *Creativity and Art: Three Roads to Surprise* (Oxford: Oxford University Press), chap. 9.

Boden, M. A. (forthcoming) 'Can Evolutionary Art Provide Radical Novelty?', in M. A. Boden and E. A. Edmonds, *Perspectives on Computer Art* (provisional title), chap. 11.

Breton, A. (1937), *L'Amour Fou* (Paris: Gallimard).

Breton, A. (1969), *Manifestoes of Surrealism* (Ann Arbor: University of Michigan Press). Trans. R. Seaver and H. R. Lane. (Includes several manifestoes, the first published in 1924.)

Cariani, P. (1992), 'Emergence and Artificial Life', in C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (eds.), *Artificial Life II* (Redwood City, CA: Addison-Wesley), 775-797.

Chalmers, D. J. (1996), *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford University Press).

Cope, D. (2001), *Virtual Music: Computer Synthesis of Musical Style* (Cambridge, Mass.: MIT Press).

Cope, D. (2006), *Computer Models of Musical Creativity* (Cambridge, Mass.: MIT Press).

Coss, R. G. (2003), 'The Role of Evolved Perceptual Biases in Art and Design', in E. Voland and K. Grammer (eds.), *Evolutionary Aesthetics* (London: Springer), 69-130.

Coss, R. G., and Moore, M. (1990), 'All that Glistens: Water Connotations in Surface Finishes', *Ecological Psychology,* 2: 367-380.

Csikszentmihalyi, M. (1999), 'Implications of a Systems Perspective for the Study of Creativity', in R. J. Sternberg (ed.), *Handbook of Creativity* (Cambridge: Cambridge University Press), 313-335.

Dehaene, S., and Naccache, L. (2001), 'Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework', *Cognition,* 79: 1-37.

Dennett, D. C. (1969), *Content and Consciousness: An Analysis of Mental Phenomena* (London: Routledge & Kegan Paul).

Dennett, D. C. (1984), *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, Mass.: MIT Press).

Dennett, D. C. (1991), *Consciousness Explained* (London: Allen Lane).

Dennett, D. C. (1995), 'The Unimagined Preposterousness of Zombies: Commentary on Moody, Flanagan, and Polger', *Journal of Consciousness Studies,* 2: 322-326.

Dretske, F. I. (1984), *Knowledge and the Flow of Information* (Oxford: Blackwell).

Dretske, F. I. (1995), *Naturalizing the Mind* (Cambridge, Mass.: MIT Press).

Edelman, G. M., and Seth, A. (2009), 'Animal Consciousness: A Synthetic Approach', *Trends in Neurosciences,* 9: 476-84.

Edelman, G. M., and Tononi, G. (2009), *Consciousness: How Matter Becomes Imagination*

(London: Allen Lane).

Fisher, E. M. (1990), *Personal Love* (London: Duckworth).

Fodor, J. A. (1995), 'Review of John McDowell's *Mind and World*', *The London Review of Books*, 17:8, April 20th, 10-11. Reprinted in J. A. Fodor, *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind* (Cambridge, Mass.: MIT Press, 1998), 3-8.

Harrison, A. (1978), *Making and Thinking: A Study of Intelligent Activities* (Hassocks, Sussex: Harvester Press).

Haugeland, J. (1978), 'The Nature and Plausibility of Cognitivism', *Behavioral and Brain Sciences*, 1: 215-226.

Hillis, W. D. (1992), 'Co-Evolving Parasites Improve Simulated Evolution as an Optimization Procedure', in C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (eds.), *Artificial Life II* (Redwood City, CA: Addison-Wesley), 313-324.

Holland, O. (ed.), *Machine Consciousness* (Exeter: Imprint Academic). Special issue of the *Journal of Consciousness Studies*, 10 (4-5).

Kacelnik, A., Chappell, J., Weir, A. A. S., and Kenward, B. (2006), 'Cognitive Adaptations for Tool-Related Behaviour in Caledonian Crows', in E. A. Wasserman and T. R. Zentall (eds.), *Comparative Cognition: Experimental Explorations of Animal Behaviour* (Oxford: Oxford University Press.), 515-528.

McDowell, J. (1994), *Mind and World* (Cambridge, Mass.: Harvard University Press).

McGinn, C. (1991), *The Problem of Consciousness* (Oxford: Basil Blackwell).

Millikan, R. G. (1984), *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Cambridge, Mass.: MIT Press).

Minsky, M. L. (1985), *The Society of Mind* (New York: Simon & Schuster).

Minsky, M. L. (2006), *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind* (New York: Simon and Schuster).

Morris, M. R. (1991), 'Why There Are No Mental Representations', *Minds and Machines*, 1: 1-30.

Morris, M. R. (1992), *The Good and the True* (Oxford: Clarendon Press).

Papineau, D. (1987), *Reality and Representation* (Oxford: Basil Blackwell).

Pattee, H. H. (1985), 'Universal Principles of Measurement and Language Functions in Evolving Systems', in J. Casti and A. Karlqvist (eds.), *Complexity, Language, and Life: Mathematical*

Approaches (Berlin: Springer-Verlag), 168-281.

Penrose, R. (1989), *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford: Oxford University Press).

Penrose, R. (1994), *Shadows of the Mind: A Search for the Missing Science of Consciousness* (Oxford: Oxford University Press).

Schaffer, S. (1994), 'Making Up Discovery', in M. A. Boden (ed.), *Dimensions of Creativity* (Cambridge, Mass.: MIT Press), 14-51.

Searle, J. R. (1992), *The Rediscovery of the Mind* (Cambridge, Mass.: MIT Press).

Sims, K. (1991), 'Artificial Evolution for Computer Graphics', *Computer Graphics*, 25 (no. 4): 319-28.

Sims, K. (1994), 'Evolving 3D-Morphology and Behavior by Competition, *Artificial Life*, 1: 353-72.

Sloman, A. (1986), 'What Sorts of Machine Can Understand the Symbols They Use?', *Proceedings of the Aristotelian Society,* Supp., 60: 61-80.

Sloman, A. (1987a), 'Motives, Mechanisms, and Emotions', *Cognition and Emotion,* 1: 217-233. Reprinted in M. A. Boden (ed.), *The Philosophy of Artificial Intelligence* (Oxford: Oxford University Press, 1990), 231-247.

Sloman, A. (1987b), 'Reference Without Causal Links', in J. B. H. du Boulay. D. Hogg and L. Steels (eds.), *Advances in Artificial Intelligence - II* (Dordrecht: North Holland), 369-381.

Sloman, A. (1993), 'The Mind as a Control System', in C. Hookway and D. Peterson (eds.), *Philosophy and the Cognitive Sciences* (Cambridge: Cambridge University Press), 69-110.

Sloman, A. (1999), 'Review of [R. Picard's] *Affective Computing', AI Magazine,* 20:1 (March), 127-133.

Sloman, A. (2000), 'Architectural Requirements for Human-like Agents Both Natural and Artificial. (What Sorts of Machines Can Love?)', in K. Dautenhahn (ed.), *Human Cognition and Social Agent Technology: Advances in Consciousness Research* (Amsterdam: John Benjamins), 163-195.

Sloman, A. (2010a), 'An Alternative to Working on Machine Consciousness', *International Journal of Machine Consciousness,* 2(1). Also available at www.cs.bham.ac.uk/research/projects/cogaff.

Sloman, A. (2010b), 'Phenomenal and Access Consciousness and the "Hard" Problem: A View from the Designer Stance', *International Journal of Machine Consciousness,* 2(1). Also available at www.cs.bham.ac.uk/research/projects/cogaff.

Sloman, A., and Chrisley, R. L. (2003), 'Virtual Machines and Consciousness', in O. Holland (ed.), *Machine Consciousness* (Exeter: Imprint Academic), 133-172.

Smith, B. C. (1996), *On the Origin of Objects* (Cambridge, Mass.: MIT Press).

Smith, T., Husbands, P., and O'Shea, M. (2002), 'Neuronal Plasticity and Temporal Adaptivity: GasNet Robot Control Networks', *Adaptive Behavior,* 10: 161-183.

Sperber, D., and Wilson, D. (1986), *Relevance: Communication and Cognition* (Oxford: Blackwell).

Todd, S. C., and Latham, W. (1992), *Evolutionary Art and Computers* (London: Academic Press).

Walker, E. H. (2000), *The Physics of Consciousness: The Quantum Mind and the Meaning of Life* (Cambridge, Mass.: Perseus).

Weir, A. A. S., Chappell, J., and Kacelnik, A. (2002), 'Shaping of Hooks in New Caledonian Crows', *Science,* 297: 981 (one page only).

Wheeler, M. W. (2005), *Reconstructing the Cognitive World: The Next Step* (Cambridge, Mass.: MIT Press).

Williamson, T. (2007) *The Philosophy of Philosophy* (Oxford: Blackwell).

Wittgenstein, L. (1953), *Philosophical Investigations,* trans. G. E. M. Anscombe (Oxford: Blackwell).

Wright, I. P. (1997), *Emotional Agents.* PhD thesis, School of Computer Science, University of Birmingham. (Available at http://www.cs.bham.ac.uk/research/cogaff/)

Wright, I. P., and Sloman, A. (1997), *MINDER1: An Implementation of a Protoemotional Agent Architecture.* Technical Report CSRP-97-1, University of Birmingham, School of Computer Science. (Available from ftp://ftp.cs.bham.ac.uk/pub/tech-reports/1997/CSRP-97-01.ps.gz)

Wright, I. P., Sloman, A., and Beaudoin, L. P. (1996), 'Towards a Design-Based Analysis of Emotional Episodes', *Philosophy, Psychiatry, and Psychology,* 3: 101-137.