CREATIVITY AS A NEUROSCIENTIFIC MYSTERY

Margaret A. Boden, University of Sussex.

Abstract: Of the three types of creativity (combinational, exploratory, and transformational), only the first has been significantly illuminated by neuroscience. And even that is not fully understood in neural terms. The other two are even more recalcitrant. This is due to the difficulty in defining the styles of thinking, and in identifying the various computational processes that are involved. One key problem is the fact that hierarchical systems cannot yet be effectively simulated in connectionist models.

I: Just what sort of mystery is this?

Many people, still under the influence of nineteenth-century Romantic views, believe that creativity is a mystery forever beyond the reach of science. They even believe, again echoing Romanticism, that it's a special faculty, confined to a tiny elite. They are wrong. In fact, creativity--that is: the ability to generate ideas/artefacts that are new, surprising, and valuable--is an aspect of human intelligence in general (Boden 2004,2010). As such, it's rooted both in our material embodiment and in our sociocultural context--and it depends on the brain. In other words, it's an unsolved puzzle for neuroscientists, not an ineluctable mystery essentially beyond their grasp.

(A point of clarification: Discussions of creativity are often bedevilled by the discussants adopting different senses of the word "new". An idea may be new to the person who has just come up with it, even though it has occurred to countless other people in the past. Let's call this P-novelty: "P" for Psychological. Alternatively, the idea may be new, so far as is known, to the whole of human history. This is H-novelty: "H" for Historical. Depending on which sense of "new" is involved, a new idea may or may not count as H-creativity. But it always exemplifies P-creativity, of which H-creativity is clearly a special case. From the psychologist's point of view, and from the neuroscientists' perspective also, the fundamental phenomenon that needs to be explained is P-creativity. Examples of H-creativity--some of which are mentioned below--may be especially interesting to us, as intellectually curious human beings. But they are scientifically relevant primarily as instances of P-creativity: their *historical* situation is not a matter for neuroscience.)

If creativity is an unsolved scientific puzzle rather than an occult and ever-enigmatic mystery, it is nevertheless a puzzle that will be very hard to solve. In common parlance, then, it's a "mystery".

The air of mystery is strengthened by the fact that introspection rarely helps. Artists, scientists, and mathematicians often report that they have no idea how they came up with their valuable new ideas. Some even use this phenomenological fact to suggest that *they* didn't come up with the novel idea at all: rather, some ultrahuman, perhaps divine, power did so. They forget, of course, that creativity is not unique in this regard: introspection doesn't tell us how we form

grammatical sentences, either, nor how we interpret photographs as depicting specific scenes. In general, much more goes on in our minds below the level of consciousness than can ever be accessed by it. (Were that not so, we'd be paralysed by information overload.) Psychology faces "introspective mystery" in all areas of mental life.

The main difficulty in solving the puzzle of creativity is not--as is also widely believed--that it is unpredictable. Creativity is indeed largely unpredictable, for a number of different reasons (Boden 2004: ch. 9). The most important reason is the enormous complexity, and idiosyncracy, of human minds, the detailed contents of which are largely unknown even to the individual concerned. Marcel Proust himself couldn't have predicted that a flood of memories would be prompted by his eating the famous madeleine. As for third-party observers, even if (which is unlikely) someone had happened to know that he used to eat madeleines as a youth at his grandmother's house, they too would have been unable to predict the host of mental associations that were triggered by his eating them again in adult life. Even if only one thought is of interest, psychological complexity may hide it from view: the very best clinical psychologist may not know whether or not Jo Bloggs will decide to commit suicide--still less just when, and how.

In one sense, this does put creativity outside the scope of science. However, that's no reason for the scientist to despair--and no reason to mark creativity off from other, notionally less mysterious, phenomena. For it's not the aim of science to predict individual events--most of which, unlike Jo Bloggs' suicide, are of no interest to us, anyway: we don't *want* physicists to be able to predict the movements of each grain of sand on the beach. (Even if the suicidal thoughts are assigned some statistical *probability,* this may not be calculated on purely scientific grounds: Meehl 1954.) Occasionally, events can be precisely predicted by science: think of a returning space-capsule, splashing into the Pacific Ocean with rescue-ships already waiting nearby. Usually, however, they cannot. Science in general isn't focussed on the prediction of particularities, even though prediction is an important aspect of experimental method. Rather, it seeks to show how events of a certain class are *possible,* and how they are related to other sorts of event, whether actual or merely conceivable (Boden 2006: 7.iii.d).

Accordingly, a neuroscientific explanation of the puzzling phenomenon of creativity would show us *how it is possible* for this still-mysterious phenomenon to occur. The common view that a science of creativity could predict every detail of creative thought, thus making human artists and scientists (and everyday punsters ...) redundant, is mistaken.

The "mystery" of creativity, as regards neuroscience, lies not in its unpredictability but in its computational variety. As outlined in Section II below, there are several different types of creativity, involving distinct sorts of information processing. A satisfactory neuroscience of creativity would have to illuminate each one of these.

"Illumination", here, means significantly more than locating the brain-areas involved. In general, a neuroscientific *explanation* of a psychological phenomenon does not merely tell us which parts of the brain, and/or which neuronal groups, are active when the phenomenon occurs. Crucially, it tells us *what the brain-cells are doing,* where this is understood not in terms of (for instance) chemical changes but in terms of the computations, or information processing, that the cells are performing (Boden 2006: ch. 14).

The computational psychologist John Mayhew, when explaining stereopsis, put it like this: "Finding a cell that recognizes one's grandmother does not tell you very much more than you started with; after all, you know you can recognize your grandmother. What is needed is an answer to how you, or a cell, or anything at all, does it. The discovery of the cell tells one what does it, but not how it can be done" (Mayhew 1983: 214).

Even if the detailed neuronal circuits involved are known, *what the circuits are doing* may be obscure. The key questions concern what information is received and/or passed on by the cell or cell-group, and how it's computed by them. Put another way, they concern "how electrical and chemical signals are used in the brain to represent and process information" (Koch and Segev 1989: 1).

The key point of this paper, then, is that we need to know *what sort of information processing* is involved in creativity, to have any hope of a neuroscientific explanation of it. And the conclusion will be that we are at present within reach of such an explanation only for one type of creativity. The others will be much more difficult nuts for the neuroscientist to crack.

II: The three types of creativity

Creativity can happen in three main ways, only one of which is typically recognized by people trying to analyse it (including those experimental psychologists who specialize in this area). Specifically, creativity may be combinational, exploratory, or transformational (Boden 2004: chaps. 3-6).

These are distinguished by the sorts of psychological process that are involved in generating the new idea. A satisfactory neuroscientific theory of creativity would need to explain how each of the three types can come about.

Combinational creativity--which is usually the only type recognized in studies/definitions of creativity--involves the generation of unfamiliar combinations of familiar ideas. In general, it gives rise to a "statistical" form of surprise, like that experienced when an outsider wins the Derby. Everyday examples of combinational creativity include visual collage (in advertisements and MTV videos, for instance); much poetic imagery; all types of analogy (verbal, visual, or musical); and the unexpected juxtapositions of ideas found in political cartoons in newspapers. Scientific examples include seeing the heart as a pump, or the atom as a solar system.

Exploratory and transformational creativity are different. Unlike the combinational variety, they're both grounded in some previously existing, and culturally accepted, structured style of thinking, or "conceptual space". Of course, combinational creativity, too, depends on a shared conceptual base--but this is, potentially, the entire range of concepts and world-knowledge in someone's mind. A conceptual space, or thinking-style, is both more limited and more tightly structured (often, hierarchically). It may be a board-game, for example (chess or Go, perhaps), or a class of chemical structures (aromatic molecules, for instance), or a particular type of music or sculpture.

In exploratory creativity, the existing stylistic rules or conventions are used to generate novel

structures (ideas or artefacts), whose possibility may or may not have been realized before the exploration took place. To the extent that it was not, the new structure will be not only satisfying but surprising. A new painting in the Impressionist style, a new benzene derivative, or a new fugue or sonnet are all examples. So is the daily generation of new sentences, fitting the grammatical rules of the language in question.

Exploratory creativity can also involve the search for, and testing of, the specific stylistic limits concerned. Just which types of structure can be generated within this space, and which cannot?

Transformational creativity is the most arresting of the three. Indeed, it leads to "impossibilist" surprise, wherein the novel idea appears to be not merely new, not even merely strange, but *impossible.* Seemingly, it simply could not have arisen--and yet it did. In such cases, the shocking new idea arose because some defining dimension of the style, or conceptual space, was altered--so that structures can now be generated which *could not* be generated before. The greater the alteration, and the more fundamental the stylistic dimension concerned, the greater the shock of impossibilist surprise.

For instance, imagine altering the rule of chess which says that pawns can't jump over other pieces: they're now allowed to do this, as knights always were. The result would be that some games of chess could now be played which were literally *impossible* before. Or consider the suggestion, new in 1865, that the benzene molecule may be a ring of carbon atoms: a topologically closed string, rather than--like all previously described molecules--an open one. *Exploratory* creativity then took over, as organic chemists mapped the space of benzene derivatives. (They later went on to ask whether the core of some ring-molecules might include five atoms rather than six, and/or atoms of elements other than carbon. Whether one chooses to call those two questions "exploratory" or "transformational" is negotiable. The important point is that they were both driven by specific features of the benzene-space that had been explored for some time.)

A comparable, and much more recent, example concerns the shocking idea that some carbon molecules may be hollow spheres. The key transformation, here, was to consider atomic bonds forming not just in one spatial dimension (as in a planar sheet of graphene), but in three. What's generally regarded as the key paper was published in 1985 (Kroto et al. 1985). It reported experimental research on carbon vapours heated to thousands of degrees, in which various multi-atom molecules (but mostly the soccer-ball C60, or Buckminster-fullerene) formed spontaneously. Subsequent *exploratory* creativity synthesized many new "fullerenes" of differing shapes and sizes. These included open-ended or closed tubes (formed when a few percent of nickel or cobalt atoms were added) that could act as molecule-carriers and electronic conductors, so providing for a host of novel technological applications. This pioneering work led to a Nobel prize eleven years later (Smalley 1996).

That work was rightly seen by the Nobel committee as "pioneering", not least because of its detail and systematicity (made possible by the team's development of laser-instrumentation for measurement). In fact, however, the central "shocking idea" had been suggested in 1970, by chemists in Japan and in the UK. But it was then considered too bizarre to be accepted (valued) by the scientific community. Moreover, a closely similar idea, envisaging the addition of impurities to a planar network of carbon atoms (and soon pointing out that the resultant hollow

molecules might carry other molecules inside them), had been published in the *New Scientist* as early as 1966--but the author had presented this as scientific fantasy rather than serious research (Jones 1966; cf. Jones 1982: 118-119). This example illustrates the difficulty, in many cases, of deciding whether a particular idea really is new, and/or really is valuable (see below).

In general (though less so in literature), transformational creativity is esteemed more highly than the other two varieties. The people whose names are recorded in the history books are usually remembered above all for changing the accepted style. Typically, the stylistic change meets initial resistance. And it often takes some time to be accepted. That's no wonder. For transformational creativity *by definition* involves the breaking/ignoring of culturally sanctioned rules.

However, novel transformations are relatively rare. All artists and scientists spend most of their working time engaged in combinational and/or exploratory creativity. That's abundantly clear when one visits a painter's retrospective exhibition, especially if the canvasses are displayed chronologically: one sees a certain style being adopted, and then explored, clarified, and tested. It may be superficially tweaked (a different palette adopted, for example). But it's only rarely that one sees a radical transformation taking place. Similarly, the list of a scientist's research papers rarely includes a transformative contribution: mostly, scientists explore the implications of some already-accepted idea. Even if that idea is itself transformative, and relatively recent, it normally prompts exploration rather than further transformation. That was so in the case of ring-molecules, as we've seen; and the case-history of the fullerenes provides further illustrations.

(Only very seldom does an individual scientist, or artist, make more than one transformative move. Picasso is an example from the arts, who pioneered several distinct styles over his lifetime. In science, the Crick-Watson team discovered both the double helix and, a few years later, the genetic code.)

The saga of the fullerenes also illustrates the fact that identifying a "creative" idea, or a scientific "discovery", is not always straightforward. Such judgments can even be affected by national rivalries, not to mention social snobbery and personal jealousies (Schaffer 1994). The identification of creativity is *never* purely scientific. For even though science can occasionally explain why we have certain values (shininess, for instance--see Boden 2006: 8.iv.c), it cannot, in principle, *justify* any value. Moreover, our values often change: different social groups/sub-groups, in differing times and places, may value very different things. Because the notion of *positive valuation* is included within the concept of creativity, the class of "creative" ideas is not a natural kind. In other words, it is not a purely scientific concept.

It follows that neuroscience could never explain the origin of creative ideas without some prior (socially based) judgements identifying *these* ideas as creative, in contrast with others that are merely new. (Even novelty isn't always easily judged, as the case-history of the fullerenes shows--see Boden 2006: 1.iii.f-g).

A final complication must be mentioned here. Namely, what we naturally think of as a "single" idea or artefact may involve more than one sort of creativity. The three forms of creativity distinguished above are analytically distinct, in that they involve different types of psychological process for generating novel ideas. But a given artwork or scientific theory can involve more than

one type. That's partly why it's generally more sensible to ask whether this or that *aspect* of the idea in question is creative, and in what way. A neuroscientific theory of creativity should be able to show how the three forms of creativity can be integrated, as well as how they can function independently.

III: What might neuroscience have to say?

There's no doubt that neuroscience could help to show how combinational creativity is possible. Indeed, it already has. Neurological studies, and computer models, of associative memory have already thrown light on the mechanisms underlying much poetic imagery.

The richness and subtlety of these associations have long been appreciated by literary scholars. The best example, here, is John Livingstone Lowes' (1930) masterly literary detective story tracing the detailed origins of Samuel Taylor Coleridge's imagery in *The Ancient Mariner* and *Kubla Khan* (Boden 2004: ch. 6). In relation to the pessimism about particularism expressed in Section I, it's worth mentioning that this author had access not only to the whole of Coleridge's eclectic library but also to his commonplace books for the eighteen months during which these poems were written, in which he had jotted down quotations that had interested him. That degree of access to the detailed contents of another person's mind is highly unusual.

However, beyond the already long-familiar idea that brains are composed of interconnected units that are somehow responsible for conceptual associations (Hartley 1749), Livingston Lowes knew nothing of the neural mechanisms involved. Today, we are in a very different position. It was known by the 1980s that certain drugs can increase or decrease the associative range of conceptual thinking, leading to more or less inclusive and/or idiosyncratic combinations respectively (e.g. Shaw et al. 1986; cf. Eysenck 1994: 224-232). And now, we have much more data, and many more neuroscientific (not least, neurocomputational) concepts, to work with.

This isn't to say that we can now come closer to literary particularism than Livingston Lowes, for instance, could. In other words, it's not to say that neuroscience could ever explain just how/why *this* idea was associated with *that* idea on a given occasion. Even if the idea in question could be neuronally located (as intentional verbs, for instance, have been located in the pSTS: Allison et al. 2000; Castelli et al. 2002; Frith and Frith 2003), the specific association that arose in some individual's mind could not be explained in detail--still less, predicted. However, we saw in Section I that particularist explanation/prediction is not the aim of science. Insofar as such particularist insights are available they are *post hoc,* not predictive, and are to be found rather in the humanities (Livingstone Lowes' discussion of *The Ancient Mariner* provides some exceptionally convincing examples).

Associative pathways, however, are not all there is to combinational creativity. There is also the tricky issue of *relevance.* Conceivably, any concept could be associated with any other, by some sufficiently tortuous neuronal path. In that sense, there's no limit to the number of "unfamiliar combinations" that are possible. But life is too short to follow only highly tortuous pathways. Even poets have to provide enough context to make their meaning communicable; and everyday speech, in general, has to be understood *immediately.* In other words, those novel combinations which we *value,* so which we regard as "creative", invariably involve

relevance--even if the relevance is not immediately apparent.

An insightful computational approach to relevance suggests that we have evolved an involuntary, and exceptionless, principle of communication (and problem-solving) based on a cost-benefit analysis, weighing effort against effect (Sperber and Wilson 1986). The more information-processing effort it would take to bear *x* in mind in the context of *y,* the more costly this would be: and high cost gives low relevance. The more implications (regarding things of interest to the individual concerned) that would follow from considering *x,* the more effective it would be: and high effectiveness gives high relevance.

The suggestion here is not (paradoxically) that we pre-compute just what effort/effect would be involved in considering a certain concept. Rather, there must be psychological mechanisms evolved for recognizing relevance. For example, our attention is naturally (sic) caught by movement, because moving things are often of interest. Similarly, even a newborn baby's attention is preferentially caught by human speech sounds. Besides being built into our sensory systems, relevance recognition is built into our memories: it's no accident, on this view, that similar and/or frequently co-occurring memories are easily accessible, being 'stored' together in scripts, schemas, and conceptual hierarchies.

Different cognitive strategies may vary in the measure of cost or benefit that they attach to a given conceptual 'distance'. Surrealists, for example, tolerate greater distances than straightforwardly 'representational' writers and painters do--hence the extreme unfamiliarity of the novel combinations found in their work. The artist's personal signature, which can affect many different aspects of a creative work (see Boden 2010), can apply here: one individual Surrealist may be even more forgiving of conceptual distance than another. Similarly, different rhetorical styles in literature involve different levels of cost and/or different types of information processing in both writer and reader: compare Charles Dickens and James Joyce, for instance. (A literary personal signature may also involve a preference for finding many sorts of relevance in certain concepts: *animals*, for the poet Ted Hughes, for example.)

This analysis of relevance implies that, *pace* symbolic computationalists such as Jerry Fodor (1983), laboured scientific inference is *not* a good model for everyday, instantaneous, understanding (Sperber and Wilson 1986: 66f.). Similarly, it rejects the GOFAI assumption that deliberate reasoning (which is needed by literary scholars and historians when puzzling over obscure texts) is required for spontaneous interpretation (op.cit., p. 75). Rather, our understanding typically depends on associative, non-logical, guessing that is constrained by what we take to be relevant.

It follows that a satisfactory neuroscientific account of combinational creativity would identify the various mechanisms evolved for judging relevance. Given that this matter is a verbal/conceptual version of the notorious frame problem (Sperber and Wilson 1996; Boden 2006: 771-5, 1003-5), that is a tall order.

With respect to the other two forms of creativity, there's more bad news. For they are significantly less amenable to neuroscience. That's true in two ways.

First, we rarely know all the constraints defining the conceptual spaces of art or science, still

less the computational processes required to explore and/or to transform them. Historians of art and musicologists spend lifetimes in attempting to make stylistic constraints explicit, and succeed only to a very limited degree. Sometimes, they even announce a given style to be unfathomable. For instance, an architectural historian specializing in Frank Lloyd Wright's work announced the style (the principle of "balance") of his Prairie Houses to be "occult" (Hitchcock 1942).

One of the advantages of computer modelling is that it can sometimes help to develop, and to test, explicit theories about such matters. So, for instance, a computerised "shape grammar" has generated every one of Lloyd Wright's forty-or-so Prairie House designs, plus many others clearly sharing the same style--without ever producing one that lacks this intuitively recognizable principle of unity (Koning and Eizenberg 1981). Moreover, this work has shown that the *fireplace* is key to the style. That is, when generating specific design-choices, changes to the location of the fireplace (or to the number of fireplaces) result in changes to most other aspects of the house.

The second type of "bad news" is that, even if we had defined the conceptual spaces concerned, and even if we knew the generative processes involved in negotiating and changing them, we wouldn't know how these are neurally embodied. We might assign them to some central cognitive workspace (e.g. Baars 1988, Changeux 2002), to be sure. And we might even be able to locate that workspace, very broadly, in the brain. But knowing just how sonnet-form, for instance, is neurally embodied, and how it is neurophysiologically accessed in generating "Shall I compare thee to a summer's day?", is way beyond the state of the art.

This is not just a difficulty in particularistic prediction, as discussed above: rather, it's a difficulty in knowing *how it is possible* for neurological mechanisms to implement sonnet-form, and to exploit it so as to generate the line in question. Similarly, explaining--in neurological terms--just how the Prairie House style can generate the Henderson house, the Martin house, or the Baker house (different examples, each named after the clients who commissioned them) is at present beyond us.

My own view is that it is likely to remain so for very many years, perhaps even forever. That's not because I agree with those philosophers (e.g. McGinn 1989, 1991) who argue that the explanation of high-level thought and consciousness is as far beyond the cognitive capacities of *Homo sapiens* as theoretical physics is beyond the capacities of squirrels and chimpanzees. I believe that position to be unnecessarily defeatist. Nevertheless, there are some fundamental problems here, which can't be solved by (theory-free) correlative brain-imaging, nor by reference, for example, to trial-and-error combinations and neural evolution (Changeux 1994).

One of these problems concerns the neural implementation of hierarchy. Most of the styles, or conceptual spaces, explored in art and science are hierarchical. The Prairie House fireplace, for instance, is key to the genre because it lies at a fundamental level in the stylistic hierarchy (the "space grammar") concerned, so that a decision about the fireplace will constrain many later decisions about other, superficially unrelated, matters. And the generation of "Shall I compare thee to a summer's day?" requires exploration of grammatical hierarchy. At present, we have no good ideas about how conceptual hierarchies are neurally embodied, nor how they can be rationally negotiated in creative thinking.

Still less do we know how transformational procedures may be embodied which can alter those hierarchies. Even domain-general transformations (such as *consider the negative* or *drop a constraint)* are a mystery. And the neural basis of the many domain-specific procedures that led from early Renaissance music (broadly: one composition, one key), through increasingly daring modulations and harmonies, to atonal music is even more elusive (Rosen 1976; Boden 2004: 71-74).

One might suggest, at this point, that computer simulation could help. And in principle, it could. However, a *neuroscientifically* plausible model is going to be connectionist rather than symbolic. Yet only symbolic models (a.k.a. GOFAI, or Good Old-Fashioned AI--Haugeland 1985: 112) are well-suited to represent hierarchy. Connectionist models, in general, are not. Despite heroic efforts in that direction, this problem has not yet been solved (Boden 2006: 12.viii). Perhaps the most impressive attempt is Harmony Theory (Smolensky et al. 1993; Smolensky and Legendre: 2006), which draws on neuroscientific knowledge. However, this was specifically developed to deal with grammatical hierarchy (syntax), and it's not clear how it could be generalized to model conceptual hierarchies such as artistic/scientific styles.

Even if it could, there would be a huge gap between harmony-theoretic modelling and the neurological reality. Most connectionist models, *especially* those intended as models of psychological (not just neurological) functions, rely on computational units which--as compared with real neurons--are too neat, too simple, too few, and too 'dry' (Boden 2006: 14.ii). In brief, the networks studied by connectionist AI are very non-neural nets.

To be sure, connectionism is becoming gradually more realistic. One recent textbook, featuring the *Leabra* software system developed by its authors, makes great efforts to integrate connectionist AI with neuroscience (O'Reilly and Munakata 2000). For example, the activation function controlling the spiking of the simulated neurons in *Leabra* is only "occasionally" drawn from mathematical connectionism (p. 42). Usually, it is based on facts about the biological machinery for producing a spike, including detailed data on ion channels, membrane potentials, conductance, leakages, and other electrical properties of nerve cells (pp. 32-48). Similarly, the basic equations used by *Leabra* when simulating high-level phenomena such as reading or conceptual memory are (usually) painstakingly drawn from detailed biophysical data. This is true, for example, of the equation used for integrating many inputs into a single neuron (see the authors' explanation of equation 2.8 on pp. 37 ff.).

The *Leabra* authors drew the line at applying this equation "at every point along the dendrites and cell body of the neuron, along with additional equations that specify how the membrane potential spreads along neighbouring points of the neuron" (p. 38). They had no wish "to implement hundreds or thousands of equations to implement a single neuron," so used an approximating equation instead. But, characteristically, they provided references to other books which did explain how to implement such detailed single-neuron simulations.

In general, the psychological models developed by O'Reilly and Munakata would have been different had the neuroscientific data been different. Their discussion of dyslexia, for instance, built not only on previous connectionist work (e.g. Plaut and Shallice 1993; Plaut et al. 1996), but also on recent clinical and neurological information (pp. 331-341). As our knowledge of the brain advances, future psychological models--they believe--will, or anyway should, be different

again. They see their book as "a ’first draft’ of a coherent framework for computational cognitive neuroscience" (p. 11).

With respect to creativity, this implies that we may hope for future connectionist models that embody specific neuroscientific data *as well as* a better understanding of the complex computational processes involved in all three types of creativity. But to hope is not to have. (And I’m not holding my breath.)

IV: Wittgenstein and neuroscience

I’ve assumed so far that it is *coherent* to aim for a neuroscientific explanation of creativity--and, for that matter, of any other psychological phenomenon. In other words, such an explanation is possible in principle, irrespective of whether it has been, or is ever likely to be, achieved. And I’ve written as though the only reason for denying this is the mysterian view that there is something essentially quasi-magical about creativity, which puts it beyond the reach of science.

However, many philosophers of mind would deny the possibility of a scientific understanding of creativity--and of any other psychological phenomenon--on very different grounds. These writers include the followers of Ludwig Wittgenstein, who suggested in his *Philosophical Investigations* (1953) that there is no level of psychological explanation between remarks about conscious phenomenology and observations about the physical mechanisms of the brain. So, for instance, Richard Rorty explicitly looked forward to "the disappearance of psychology as a discipline distinct from neurology" (1979: 121).

Wittgensteinians in general reject psychological explanations posed at the sub-personal level, so criticize those *neuroscientific* theories which define brain-processes cognitively (or computationally), rather than purely neurologically. They accuse neuroscientists of incoherence due to the "mereological fallacy", which is to attribute to a part of a system some predicate which is properly attributed only to the whole (Bennett and Hacker 2003). In this context, the "system" in question is the whole person, the "parts" are the brain (or parts thereof), and the "predicates" are psychological terms such as knowledge, memory, belief, reasoning, choice--and, of course, creativity.

On this view, there is absolutely no hope of a naturalistic psychology. Insofar as psychology exists as a scientific discipline it is said to be a hermeneutic, not a natural science (cf. McDowell 1994; Harre 2002). So neuroscience could never *replace* psychology, in the sense of substituting for it. At most, a (non-cognitivist) neuroscience could compensate for the lack of a cognitivist (sub-personal) psychology.

This rejection of naturalism in psychology reflects a deep divide in western philosophy, which we can’t go into here (but see Boden 2006: 16.vi-viii). A few neuroscientists (such as followers of Humberto Maturana and Francisco Varela: 1980) lie on the anti-naturalist side of the divide. But the vast majority do not. Moreover, neuroscience itself has become increasingly cognitivist--indeed, computational--since the 1950s (Boden 2006: ch. 14). Information-processes and computational mechanisms are now considered crucial in many neuroscientific explanations, from studies of vision to the higher thought processes. And this paper has argued that the

computational level of theorizing is crucial in explaining creativity, too.

So although Wittgenstein might seem at first sight to be the neuroscientist's friend, perhaps he is not such a good friend after all. Not a false friend, to be sure (for that would involve insincerity or betrayal). But, in my view, a mistaken one.

V: Conclusion

Nothing that's been said above suggests that there can never be a neuroscience of creativity. Indeed, a neuroscience of *combinational* creativity is arguably within sight--if not yet within reach.

It's not yet in reach, partly because--as explained in Section III--there are challenging problems concerning how we make judgments of *relevance* when engaging in, or appreciating, combinational creativity. A neuroscientific explanation of that is not within sight. Moreover, given that this is a verbal/conceptual version of the notorious frame problem (Sperber and Wilson 1996; Boden 2006: 771-5, 1003-5), it is a tall order.

Further reasons why a neuroscience of creativity is not within reach involve hierarchy, as we've seen. Clearly, it must be possible, somehow, for hierarchy--and all other aspects of symbolic thinking--to be implemented in (broadly) connectionist systems. After all, the human brain is such a system. However, we need to understand, much better than we do at present, how a basically connectionist system can emulate a symbolic one (how connectionism can emulate a von Neumann machine).

In addition to highly general questions such as that one, we need to focus on the specific structure of, and the generative processes within, the myriad conceptual spaces underlying science and art. For neither exploratory nor transformational creativity can be properly understood without taking those computational features into account.

References:

Allison, T., Puce, A., and McCarthy, G. (2000), 'Social Perception from Visual Cues: Role of the STS Region', *Trends in Cognitive Sciences* 4(7): 267-278.

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

Bennett, M. R., and Hacker, P. M. S. (2003), *Philosophical Foundations of Neuroscience* (Oxford: Blackwell).

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), ’Personal Signtaures in Art’, in M. A. Boden, *Creativity and Art: Three Roads to Surprise* (Oxford: Oxford University Press), 92-124.

Castelli, F., Frith, C. D., Happe, F., and Frith, U. (2002), ’Autism, Asperger Syndrome, and Brain Mechanisms for the Attribution of Mental States to Animated Shapes’, *Brain,* 125: 1839-1849.

Changeux, J.-P. (1994), ’Creative Processes: Art and Neuroscience’, *Leonardo, 27(3): 189-201.*

Changeux, J.-P. (2002), *The Physiology of Truth: Neuroscience and Human Knowledge,* trans, M. B. DeBevoise (Cambridge, Mass.: Harvard University Press).

Eysenck, H. J. (1994), ’The Measurement of Creativity’, in M. A. Boden (ed.), *Dimensions of Creativity* (Cambridge, Mass.: MIT Press), 199-242.

Fodor, J. A. (1983), *The Modularity of Mind: An Essay in Faculty Psychology* (Cambridge, Mass.: MIT Press).

Frith, U., and Frith, C. D. (2003), "Development and Neurophysiology of Mentalizing’, *Phil. Trans. Royal Society of London B* 358: 459-473.

Harre, R. M. (2002) *Cognitive Science: A Philosophical Introduction* (London: Sage).

Hartley, D. (1749), *Observations on Man: His Frame, His Duty, and His Expectations.* London. (Facsimile reproduction ed. T. L. Huguelet, reprinted Gainesville, Florida: Scholars’ Facsimiles and Reprints, 1966.)

Haugeland, J. (1985), *Artificial Intelligence: The Very Idea* (Cambridge, Mass.: MIT Press).

Hitchcock, H. R. (1942), *In the Nature of Materials: The Buildings of Frank Lloyd Wright, 1887-1941* (New York: Meredith Press).

Jones, D. E. H. [as *Daedalus]* (1966), ’Note in Ariadne column’, *New Scientist,* 32 (3rd November): 245.

Jones, D. E. H. (1982), *The Inventions of Daedalus* (Oxford: W. H. Freeman).

Koch, C., and Segev, I. (eds.) (1989), *Methods in Neuronal Modeling: From Synapses to Networks* (Cambridge, Mass.: MIT Press).

Koning, H., and Eizenberg, J. (1981), ’The Language of the Prairie: Frank Lloyd Wright’s Prairie Houses’, *Environment and Planning, B,* 8: 295-323.

Kroto, H. W., Heath, J. R., O’Brien, S. C., Curl, R. F., and Smalley, R. E. (1985), ’C60: Buckminsterfullerene’, *Nature,* 318(6042): 162-163.

Livingston Lowes, J. (1930), *The Road to Xanadu: A Study in the Ways of the Imagination* (London: Houghton). Revd. edn., 1951.

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

McGinn, C. (1989), 'Can We Solve the Mind-Body Problem?', *Mind*, 98: 349-366.

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

Maturana, H. R., and Varela, F. J. (1980), *Autopoiesis and Cognition: The Realization of the Living* (Boston: Reidel). First published in Spanish, 1972.

Mayhew, J. E. W. (1983), 'Stereopsis', in O. J. Bradick and A. C. Sleigh (eds.), *Physical and Biological Processing of Images* (New York: Springer-Verlag), 204-216.

Meehl, P. E. (1954), *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (Minneapolis: University of Minnesota Press).

O'Reilly, R. C., and Munakata, Y. (2000), *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (Cambridge, Mass.: MIT Press).

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996), 'Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains', *Psychological Review,* 103: 56-115.

Plaut, D., and Shallice, T. (1993), 'Deep Dyslexia: A Case Study of Connectionist Neuropsychology', *Cognitive Neuropsychology,* 10: 377-500.

Rorty, R. (1979), *Philosophy and the Mirror of Nature* (Princeton, N.J.: Princeton University Press).

Rosen, C. (1976), *Schoenberg* (Glasgow: Collins).

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

Shaw, E. D., Mann, J. J., and Stokes, P. E. (1986), 'Effects of Lithium Carbonate on Creativity in Bipolar Outpatients', *American Journal of Psychiatry,* 143: 1166-1169.

Smalley, R. E. (1996), 'Discovering the Fullerenes', *Reviews of Modern Physics,* 69(3): 723-730. (Nobel Prize acceptance speech.)

Smolensky, P., and Legendre, G. (2006), *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar,* 2 vols. (Cambridge, Mass.: MIT Press).

Smolensky, P., Legnendre, G., and Miyata, Y. (1993), 'Integrating Connectionist and Symbolic Computation for the Theory of Language,' *Current Science* 64: 381-391.

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

Sperber, D., and Wilson, D. (1996), 'Fodor's Frame Problem and Relevance Theory', *Behavioral and Brain Sciences*, 19: 530-532.

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