Steven Shaviro

discognition

Repeater
Discognition
Discognition
Steven Shaviro
CONTENTS

Preface
Introduction
ONE Thinking Like a Philosopher
TWO Thinking Like a Computer
THREE Thinking Like an Avatar
FOUR Thinking Like a Human Being
FIVE Thinking Like a Killer
SIX Thinking Like an Alien
SEVEN Thinking Like a Slime Mold
AFTERWORD Twenty-Two Theses on Nature
References
Works Cited
Preface


As always, this writing is for my daughters, Adah and Roxanne.
Introduction

What is consciousness? How does subjective experience occur? Which entities are conscious? Or, to put things as particularly as possible: what is it like to be a bat? as – Thomas Nagel famously asked. For that matter, what is it like to be a dog, a robot, or a tree – or even a human being? Is it like anything at all to be a rock, or a star, or a neutrino? How do we explain the very fact of being aware? What does it really mean to be conscious, to think, to feel, or to know? And what is the difference – if any – between thinking, feeling, being aware, and knowing? Such questions might seem to have obvious answers – until we actually try to answer them. Then we discover that we don’t have a clue, and that these questions have never come close to being plausibly answered. Still today, there is no consensus whatsoever upon any of these topics: neither among scientists and philosophers, nor among the general public. We are clearly sentient, and yet we do not know what sentience is, how it can exist, or what it means.

Whenever I come across such intractable problems, my impulse is always to turn to science fiction. Perhaps we will be able to imagine what we are unable to know. Science fiction is a special kind of literature – or better, paraliterature, as Samuel R. Delany calls it – that operates through speculation and extrapolation, and that takes place (conceptually, if not grammatically) in the future tense. It is a kind of thought experiment, a way of entertaining odd ideas, and of asking off-the-wall what if? questions. But instead of approaching its issues abstractly, as philosophy does, or breaking them down into empirically testable propositions, as physical science does, science fiction embodies these issues in characters and narratives. By telling stories, it asks questions about all sorts of things: consciousness and cognition, the future, extreme possibilities, nonhuman otherness, and especially the deep consequences – the powers and limitations – of both our ideologies and our technologies.

The method of science fiction is emotional and situational, rather than rational and universalizing. Philosophical argumentation and scientific
experimentation both endeavor to prove and to ground their assertions, however counterintuitive these may seem to be at first glance. Science fiction also proposes counterintuitive scenarios; but its effort is rather to work through the weirdest and most extreme ramifications of these scenarios, and to imagine what it would be like if they were true. Where philosophy is foundational, science fiction is pragmatic and exploratory. And where physical science seeks to settle upon predictable and repeatable results, science fiction seeks to unsettle and singularize these results, and to provide us with unrepeateable histories. Science fiction does not ever actually prove anything; but its scenarios may well suggest new lines of inquiry that analytic reasoning and inductive generalization would never stumble upon by themselves.

In Discognition, I look at a series of science fiction narratives in order to raise questions about consciousness and thought – or better, about sentience. I prefer this latter term, because it does not presuppose that mental processes and experiences are rational, nor even that they are necessarily conscious. When certain philosophers elevate human “sapience” over mere animal “sentience”, they are indulging in dubious feats of self-congratulation. For in fact, there is far more of an evolutionary continuity than a sharp distinction between the way that my dog thinks, and the way that I think. I have many unique qualities of mind that he can never hope to possess; but the inverse of this is also true. Understanding and intelligence (which Robert Brandom lists as the characteristics of sapience) are in fact deeply rooted in such features of sentience as sensory awareness, reality testing, irritability, and arousal. The difference is one of degree, rather than one of kind.

Brandom is therefore wrong to scornfully dismiss what he calls the “merely sentient” condition of animals. My dog may not be able to “offer and inquire after reasons”, as Sellars and Brandom would wish – just as he cannot figure out how to extricate himself when he gets tangled up in his leash. Nonetheless, he exhibits a wide range of moods and feelings. He is quite good at posing and pursuing many sorts of complicated goals. And he is highly skilled at expressing his desires, in ways that I am able to understand; and at comprehending – and responding flexibly to – my own moods and desires. Thinking is a far more common and widely distributed process than we are sometimes willing to recognize.
The narratives that I discuss in this book offer us speculation – fictions and fabulations – about sentience. There is something oddly recursive about this, since sentience itself is arguably a matter of generating (or being able to generate) fictions and fabulations. We ought to resist the all-too-common equation of sentience with cognition. We often find this assumption taken for granted in contemporary philosophy of mind, as well as in neurobiological research. But mental functioning and subjective experience need not themselves be cognitive – even though cognition seems impossible without them. Sentience, whether in human beings, in animals, in other sorts of organisms, or in artificial entities, is less a matter of cognition than it is one of what I have ventured to call discogniton. I use this neologism to designate something that disrupts cognition, exceeds the limits of cognition, but also subtends cognition. My working assumption is that fictions and fabulations are basic modes of sentience; and that cognition per se is derived from them and cannot exist without them.

Fictions and fabulations are often contrasted, or opposed, to scientific methods of understanding the world. But in fact, there are powerful resonances between them; they are both processes of speculative extrapolation. In other words, constructing and testing scientific hypotheses is not entirely different from constructing fictions and fabulations, and then testing to see whether they work or not, and what consequences follow from them. For science is far more than just a passive process of discovery, or a compiling of facts that are simply “out there”. Rather, science must actively approach things and processes in the world. This is the reason for making hypotheses. Science needs to solicit and elicit phenomena that would not disclose themselves to us otherwise. It must somehow compel these phenomena to respond to our questions, by giving us full and consistent answers. All this is necessary, precisely because things in the world are not cut to our measure. They have no reason to conform to our presuppositions, or to fit into any categories that we seek to impose.

The modern empirical scientific method is sometimes described as a process of “torturing nature to reveal her secrets” – a phrase often wrongly attributed to Francis Bacon. Philosophers of science also like to quote Isaac Newton’s Hypotheses non fingo (“I feign no hypotheses”). But a much better account of actual scientific practice is the one proposed by Bruno Latour and Isabelle Stengers, who say that scientists work by negotiating with nonhuman entities, and by entering into alliances with them. Scientists
do not get very far by treating the things they are interested in as mute and inert objects to be dissected. They do much better when they are somehow able to collaborate with the very entities that they seek to observe and explain.

Alfred North Whitehead, a major inspiration for both Latour and Stengers, notes that if the “rigid… Baconian method of induction” had been “consistently pursued”, it “would have left science where it found it”. Nothing new would ever have been discovered. The same can be said for Newton’s claim of making no hypotheses. Whitehead insists that science needs not just empirical observation and induction, but also “the play of a free imagination, controlled by the requirements of coherence and logic”. That is to say, a certain degree of speculation is always necessary in scientific research. This speculation has to be “controlled” in some manner; it cannot be altogether arbitrary and unbounded. But without speculation, science is caught in a rut. It cannot stretch beyond the given, immediate facts, in order to provide a plausible explanation for these facts.

The speculative process described by Whitehead is roughly similar to what Charles Sanders Peirce calls abduction. For Peirce, abduction stands in contrast to – and supplements – both deduction and induction. Deduction starts with conditions that are already given, and traces out a chain of logical consequences for those conditions. Induction, for its part, generalizes on the basis of an already given set of particular observations. According to Peirce, neither deduction nor induction can actually suggest anything new. Abduction, in contrast, makes a sort of leap into novelty. It shifts register: suggesting a higher-order explanation for the circumstances with which it is concerned, or positing a possible cause for the effects in view. Science is often praised for having – as other human disciplines do not – an intrinsic self-correcting mechanism. But without first engaging in abduction or speculation, science would never come up with any material to confirm or deny, or to self-correct.

Because it requires flights of speculation, as well as because it requires collaboration among many separate entities, science can never be purely human, nor purely rational. This is why efforts to place science on a pedestal, radically separating it from other forms of thought and endeavor, are so deeply mistaken. Empirical science and rational discourse are largely continuous with other ways of feeling, understanding, and engaging with
the world. These include art, myth, religion, and narrative, together with the nonhuman modes of inference exhibited by other sorts of organisms.

We should therefore always be alert to the deep biological roots of scientific experimentation and discovery. As Björn Brembs points out, there has recently been a major change of paradigm in neuroscience: a “dramatic shift in perspectives from input/output to output/input”. We can no longer be satisfied with the old stimulus/response model, according to which animals (and other organisms) passively respond to prior, incoming stimuli, and learn by means of conditioning (or associations among these stimuli). For this is only one part of the story. In addition, and much more importantly, biological entities are active reality-testers. They are always busy “probing the environment with ongoing, variable actions first and evaluating sensory feedback later (i.e., the inverse of stimulus response)”. Rather than just responding to stimuli, they exhibit ongoing activity that is self-generated, and only secondarily modulated by stimuli. Output tends to come first. Organisms engage with their surroundings with spontaneous actions, rather than just waiting for and responding to sensory inputs.

For instance, fruit flies (the special focus of Brembs’ own research) only have tiny brains; but they actively compare the actual results of their reality-testing with what can only be called their prior expectations. They also engage in spontaneous (non-deterministic and unpredictable) actions, so that their behavior “is notoriously variable, even under identical sensory conditions”. The same applies, not just to animals with neurons and brains, but also to non-animal forms of life, like trees, bacteria, and slime molds. That is to say, living organisms are continually engaged, in their own particular ways, in processes of speculative extrapolation and experimentation. When scientists perform experiments and develop theories, actively soliciting responses from the world, they are fundamentally doing the same thing as fruit flies and slime molds – albeit in a far more sophisticated manner, and on a more reflexive meta-level.

Among human beings, speculative extrapolation is not only the method of science. It is also what art in general does – and what science fiction does in particular. As the philosopher Eric Schwitzgebel puts it:

Increasingly, I think the greatest science fiction writers are also philosophers. Exploring the limits of technological possibility
inevitably involves confronting the central issues of metaphysics, epistemology, and human value.

In this book, I seek to explore the potentials and implications of sentience by turning to fictions and fabulations – and in particular to written science fiction narratives. I explore texts that are set mostly in the near future, and trace out the potential implications of already-existing technologies and research programs in the science and philosophy of mind. Some of these stories can be described as reductionist and eliminativist, in the sense that they seek to demystify and discredit our common sense assumptions about how our minds work. Others might be described as expansive, in that they seek to show that phenomenal consciousness is irreducible, and more widely spread than we sometimes imagine. Some of the narratives deal with human intelligence and consciousness in particular; others propose radically alien sorts of mentality. In all cases, I seek to follow, and extrapolate from, the suggestions expressed by the narratives themselves – rather than viewing them with suspicion, or working to critique them.

More specifically, the hypothesis, or speculative wager, behind this book is that science fiction narratives can help us step beyond the overly limited cognitivist assumptions of recent research both in the philosophy of mind and in the science of neurobiology. This is because narrative fictions nearly always extend beyond cognition. They are about connecting how and what we know to *how we feel*, and to *how we might act*: to *what is it like?* in short. Even the most reductionist SF stories still work, not just to explain, but also to entangle us within their grim scenarios. In this sense, works of art are forms of – or occasions for – *rehearsal*, as Morse Peckham argued long ago. With their extrapolations, they allow us to respond vicariously to situations that might be extremely dangerous and painful, were they actually to exist. Art readies us for evaluation and action under conditions of uncertainty. In the aesthetic register, Peckham says, “responses are redundantly maintained in situations in which nothing is at stake”. This is precisely what allows narrative (and other forms of art) to explore extreme possibilities.

Psychoanalysis and cognitive science both tell us – albeit for vastly different reasons – that consciousness is only a very narrow and specialized part of mental activity. Most thinking takes place nonconsciously, outside of our attention or awareness. Even more of our thinking slips away – it
cannot be retained in memory, or in the form of concepts. Fictions and fabulations can provide us with a sort of feed forward to use a phrase of Mark Hansen’s – of those mental processes that are not available to introspection. Hansen emphasizes the (quite science-fictional) way that computational microsensors are now able “to stand in for consciousness, to take the place of sense perception in the operations of registering sensory data”. Things beneath or beyond the reach of phenomenal perception are thus made accessible to us, albeit belatedly and indirectly. I want to suggest that fictions and fabulations, whether articulated by human beings or by other entities, are also forms of indirect, nonphenomenological access to nonconscious forms of sentience.

Through fictions and fabulations, we learn that there is more to thought than consciousness. But there is also more to thought than the nonconscious computations of which cognitive science speaks. Before it is cognitive, let alone conscious, thought is primordially an affective and aesthetic phenomenon. This is best grasped as a process of what Alfred North Whitehead calls “feeling”. Whitehead uses this word, he says, as “a mere technical term” in order to designate “that functioning through which the concrescent actuality appropriates the datum so as to make it its own”. What this means, in more familiar language, is that every entity becomes what it is by “appropriating” what is left behind by other entities that precede it. Most crucially, an entity perpetuates itself by appropriating its own prior states of existence. But an entity also appropriates other entities in its surroundings. It picks up whatever it encounters: whatever affects it, or provides conditions or resources for its own continued existence.

This primordial act of feeling, or appropriation, happens before I know it, and often without my ever becoming aware of it. I can breathe without having to know anything about oxygen. Feeling, as Whitehead describes it, comes about prior to anything like understanding (in the Kantian sense), or cognition (in the current psychological and analytic-philosophical sense) or intentionality (in the phenomenological sense). Rather, Whiteheadian feeling is closer to Spinoza’s notion of affection (affectio), and to William James’ theory of emotion. Embodied response precedes, and does not require, intellectual apprehension.

In other words, feeling is something that happens without, or before, concepts. Modern philosophy is generally uncomfortable about this prospect. Think, for instance, of Kant’s dictum that “thoughts without
content are empty, intuitions without concepts are blind”; of Merleau-Ponty’s insistence that “unreflective experience” must itself be reflected upon, and that such reflection “cannot be unaware of itself as an event”; and of Sellars’ attack on the “myth of the given”. All of these philosophers insist that there is no such thing as raw, unmediated experience. Our perceptions and emotions are always already conceptualized. Of course these arguments are in their own terms impregnable; if I want to insist upon a “feeling” that is prior to these modes of conceptualization and self-reflection, then I cannot go on to conceptualize it. I cannot assume its solidity as an idea, or as a point of presence. I must regard feelings, and characterize them, as fugitive and ungraspable; and perhaps also as nonfunctional, or even dysfunctional.

This means, in Kantian terms, that “feeling” is a matter for aesthetics, rather than for the empirical understanding. Despite his strictures against “intuitions without concepts” in the First Critique, Kant nonetheless writes in the Third Critique of “aesthetic ideas”, which he defines as “inner intuitions” that are so powerful that “no concept can be fully adequate to them”. In phenomenological terms, we may say that feeling comes before, and falls short of, any sort of intentionality, or even of Merleau-Ponty’s reversibility. In cognitivist terms, finally, feeling has something to do with what Thomas Metzinger calls Raffman qualia. A sensation of this sort is “available for attention and online motor control, but it is not available for cognition… it evades cognitive access in principle. It is nonconceptual content”.

In his recent book Plant-Thinking, Michael Marder credits plants with “non-conscious intentionality”. He means “intentionality” in the phenomenological sense: the idea that thought is of or about something. In this book, I argue pretty much the reverse: that living organisms, beyond and beneath their cognitive accomplishments, exhibit something like nonintentional sentience. Beneath intentionality, or before thought is about anything, there is a thinking process – an it thinks – that is nontransitive, without an object. When it thinks, it feels something; but it does not have any conception or representation of what it is that it feels. As Marder rightly points out, plants do not have anything like a unified or centered self. There is no “I” to a plant, no subject. But for this very reason, there is nothing – as far as a plant is concerned – like an intentional object either. My formulation is not an absolute reversal of Marder’s, because I do not equate
sentience with consciousness. I think that Whitehead is right in speaking of the relative rarity of consciousness, and in suggesting that most occasions of feeling are nonconscious. Plants are indeed sentient, as recent research has convincingly shown. But this does not necessarily mean that they are conscious. Plants feel, in Whitehead’s sense; they encounter the world. But they do not do so in any manner with which we are acquainted.

In *Discognition*, I look at science fiction narratives – fictions and fabulations – that consider unusual forms of sentience, both in human beings and in other entities. The first chapter, “Thinking Like A Philosopher”, is not about a science fictional text *per se*, but rather about a counterfactual narrative – the story of Mary – that has become the focus of much speculation and argumentation among philosophers of mind. The second chapter, “Thinking Like A Computer”, discusses Maureen McHugh’s short story “The Kingdom of the Blind”, which contemplates the possibility of spontaneously arising machine sentience, or artificial intelligence. The third chapter, “Thinking Like An Avatar”, looks at Ted Chiang’s dramatization of the issues surrounding artificial intelligence in his novella “The Lifecycle of Software Objects”. The fourth chapter, “Thinking Like A Human Being”, considers R. Scott Bakker’s chillingly eliminativist view of human cognition, as expressed in his novel *Neuropath*. The fifth chapter, “Thinking Like A Killer”, looks at Michael Swanwick’s short story “Wild Minds”, which considers the development of a process to “optimize” consciousness. “Wild Minds” was written long before *Neuropath*, but it almost seems like a deliberate rejoinder to that novel. The sixth chapter, “Thinking Like An Alien”, examines Peter Watts’ First Contact novel *Blindsight*. This book raises questions about the very nature of consciousness by imagining radical, posthuman mind alterations alongside a truly alien sort of intelligence. The seventh chapter, “Thinking Like A Slime Mold”, considers the strange cognitive powers of the plasmodial slime mold *Physarum polycephalum*, an actually existing organism that seems like a creature of science fiction. In lieu of a conclusion, the Afterword presents a number of speculative theses about Nature. These do not follow in any direct sense from the readings of science fiction in the previous chapters, but they provide a broad framework within which all the arguments of *Discognition* may be placed.
Thinking Like a Philosopher

There’s a famous story in the philosophy of mind: the story of Mary. It goes something like this. Mary is the world’s greatest neuroscientist. She knows everything there is to know about the physical world, and about how our brains work to perceive and interpret the world. In particular, Mary knows everything there is to know about color and color vision: from the physics of light, to the structure of the eye and the nervous system in human beings and other organisms, to the ways that our brains recognize and distinguish particular colors, to the evolutionary origins of color vision, to the functions served in our minds by the apprehension of color, and the ways that our moods are affected by seeing one color or another. In short, Mary has all the “physical information” about color: she knows every material and scientific fact that there is to know. The science of color is complete: nothing about it remains undiscovered.

But there’s a catch to the story. Despite her exhaustive knowledge, Mary herself has never perceived any color at all. She has lived for her entire life in a room that is entirely black and white. She has read black-and-white textbooks, and watched black-and-white videos. And so she knows that the sky is blue, that grass is green, and that roses are red. But she has never actually seen the sky, the grass, or a rose. She has only read about them, or viewed black-and-white photos and videos of them.

The question is: what happens when Mary finally leaves her black-and-white room, goes outside, and sees a red rose for the very first time? What does it mean for her to feel, for herself, what she has previously only known about? What is it like for her to perceive the color red? Does the phenomenal experience of redness add anything to her store of knowledge about the color, and about how people respond to it? Does Mary learn something that she didn’t know before?

The story of Mary is an exemplary science fiction narrative. It imagines a fantastic scenario: one that goes beyond our present scientific knowledge and technological abilities, but that – at least in principle – remains within
the bounds of scientific possibility. This doesn’t mean that I would actually expect such an experiment to be carried out. My mind boggles at the thought of how sadistic it would be to submit a human being to imprisonment and deprivation of this sort, all in the name of posing a philosophical puzzle. (Not to mention — as Moira Gatens suggested to me, when I gave an earlier version of this chapter as a talk — the problem of keeping Mary’s existence color-free even when she had her period.) I also wonder about how expensive Mary’s confinement would be, and who would pay for it. Perhaps Mary’s life in the black-and-white room could be broadcast to the public as reality television. Think of the way Jim Carrey’s personal life becomes a spectacle for millions in the 1998 movie *The Truman Show*.

Any thorough, materially grounded consideration of the story of Mary would need to take issues like these into account. Nonetheless, the story of Mary works as science fiction even when we bracket such pragmatic concerns, because of the way that it invites speculation. It asks us to consider what visual perceptual experience is like, and what we can learn from it. More specifically, it poses the question (which has vexed philosophers for hundreds of years, and which seems nearly impossible to test scientifically) of how private inner experience — like the sensation of seeing the color red — relates to public, outward linguistic expression. Does the latter adequately describe or correspond to the former? In what sense can the former be said to exist, without the latter?

The story of Mary was not actually published as science fiction, however. It was invented by the philosopher Frank Jackson, in the course of his 1986 paper, “Epiphenomenal Qualia”. Jackson’s account exemplifies the way that analytic philosophers often construct bizarre scenarios, in order to test the extreme consequences of their arguments. I think of this strategy as a form of science fictional speculation. As the philosopher of mind Eric Schwitzgebel puts it:

A good science fiction writer can open your mind up to possibilities that you might not have considered before, can break you out of your culturally-given shell of presuppositions about how the world must be. I especially like science fiction that explores possibilities around amplification of our cognitive powers and what this means for our sense of personhood and our values.
Most contemporary philosophers are not as forthcoming as Schwitzgebel about how their own work uses the methods of overt science fictional fabulation. But the similarity is evident. In addition to the story of Mary, philosophers have imagined such scenarios as the following:

1. Inverted visible spectra, so that everything that one person experiences as “red” is experienced by another person as “green”.
2. Worlds in which water has exactly the same properties as it has for us, but in which the chemical formula for this “water” is not $\text{H}_2\text{O}$.
3. Brains in vats being fed simulated experiences by direct neurochemical stimulation.
4. Zombies who are physically indistinguishable from actual people but who lack any sort of consciousness or inner experience.

The lineage of this sort of speculative-fiction-as-philosophy extends back at least to Descartes’ hypothesis of an Evil Demon. In more contemporary terms, these speculative scenarios are not far from the ones imagined in the novels of Philip K. Dick, or in movies like *The Matrix*. Analytic philosophers, no less than science fiction authors, engage in a systematic practice of speculative extrapolation. Weird and extreme scenarios can challenge our everyday assumptions, and push actually existing conditions to their most far-reaching possibilities. In reading the story of Mary as a science fiction narrative, I will consider both the powers and the limitations of such fabulations as a form of philosophical inquiry.

One reason for proposing speculative scenarios is to challenge our common intuitions. These would seem to suggest that Mary does, at the very least, encounter something new when she leaves her room. Redness is an example of what philosophers call *qualia*: phenomenal sensations, or “raw feels”, that seem to make up the very fabric of our mental experience. And the qualia of color vision, in particular, are precisely what Mary is missing inside her black-and-white room. Until she actually sees a red object, therefore, Mary does not know, from the inside, *what it is like* to experience redness. But how does this square with the supposition that, while still stuck inside the room, she *already* knows everything that there is to know physically, materially, and scientifically – about color?
Frank Jackson calls himself a “qualia freak”, and he initially poses the story of Mary in these terms. He argues that “there are certain features of the bodily sensations especially, but also of certain perceptual experiences, which no amount of purely physical information includes”. For the “information” we get by *experiencing* something is quite different from the “physical information” that allows us to say that we know about it, or understand it. No amount of objective physical information can tell us, Jackson claims, “about the hurtfulness of pains, the itchiness of itches, pangs of jealousy, or about the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky”. Jackson concludes that physicalism – the doctrine that everything in the world is physical or material – must be wrong. For any description of the world in exclusively physical terms excludes qualia, and therefore is radically incomplete.

In the decades since Jackson first published the story of Mary, it has been the subject of scores of articles by analytic philosophers. Nearly all of these thinkers have responded to Jackson’s challenge by seeking to account for qualia and phenomenal experience in a way that does not lead to his anti-physicalist conclusion. If physicalism is true, and everything in the world is composed of material stuff, then there must be some flaw in the logic of Jackson’s argument. Even Jackson himself has come to embrace this position. He now says, rather disparagingly, that no mere “epistemological claim”, such as he made in his story about Mary, can get in the way of the basic metaphysical truth of physicalism. We may not know how to “deduce” the physical basis of qualitative “psychological states” from the information that we have, but it does not follow that these states are therefore devoid of any such physical basis at all.

But there’s one serious problem with all these philosophical discussions. Even though nearly everyone agrees that there is something fundamentally wrong with Jackson’s initial claims, nobody can agree as to *just where* the mistake lies. Every philosopher has a different account of what is wrong. Daniel Dennett, for instance, argues that the whole story of Mary “is a bad thought experiment, an intuition pump that actually encourages us to misunderstand its premises”. If Mary *really* knew all the physical facts about color, Dennett says, then she would already know what it is like to have the sensation of seeing red. She would not learn anything new when she actually saw a red object for the first time.
Dennett even composes his own weird science fiction scenario in order to reinforce his point. He proposes that, when Mary leaves her room, somebody tries to trick her by showing her a banana that is painted blue. The trickster hopes that, since Mary knows from her readings that bananas are supposed to be yellow, she will mistake the qualitative feel of blue for that of yellow. But Dennett insists that Mary cannot be fooled, because she already knows, “in exquisite detail, exactly what physical impression a yellow object or a blue object (or a green object, etc) would make on [her] nervous system”.

For Dennett, if Mary really knows everything there is to know about color, then she must already know what it feels like to encounter this or that color. The story misleads us when it suggests that there could be an experience outside of discursive knowledge. Dennett denies that there can be anything like an ineffable experience of qualia; for him, “what is it like?” is not a meaningful question. Indeed, Dennett insists that “qualia” do not exist in the first place. At the very least, they do not have the special qualities that Jackson – following common sense – attributes to them. Dennett argues that so-called “qualia” are nothing more than “mere complexes of mechanically accomplished dispositions to react” to various stimuli. There is no mystery about first-person phenomenal experience, he says, because there is nothing more to it than such mechanistic habits.

The late David Lewis, in contrast to Dennett, accepts that Mary does in fact learn something new when she exits the black-and-white room. But he denies that what she learns is a new fact, beyond the physical facts she knew already. Mary does not gain any new propositional knowledge, Lewis says. Rather, she acquires something like “know-how”, or the instrumental ability “to remember and imagine and recognize” the color red. “Knowing-how” to do something is not the same as “knowing-that” something is the case. In this way, the novelty of Mary’s actual experience of colors is admitted, but the physicalist claim that physical facts are the only facts is preserved.

Lewis, unlike Dennett, concedes that Mary could not have acquired her know-how about the color red simply by studying all the facts about color from inside her black-and-white room. But he still insists that there is nothing special about inner, subjective experience. Indulging in his own science fictional speculation, Lewis suggests that Mary could also get her pragmatic know-how about color in other ways. For instance, she might
acquire the ability to recognize red through “precise neurosurgery, very far beyond the limits of present-day technique”. Such surgery would implant in her brain the very same neurochemical configurations, and therefore the same instrumental abilities, that are produced within her when she first learns to recognize the color red. Like a character in a Philip K. Dick story, Mary would remember what it is like to experience the color red, despite never actually having had such an experience.

In contrast to both Dennett and Lewis, Michael Tye argues that, when Mary leaves her room, she actually does learn something new; and that what she learns is not just a pragmatic ability like Lewis’ “know-how”. Rather, according to Tye, when Mary leaves her room she develops a new “phenomenal concept” of red. This “phenomenal concept” is the knowledge of “what it is like” to experience red; it plays the “functional role” of allowing Mary to “discriminat[e] the experience of red from other color experiences in a direct and immediate manner via introspection”. Mary thus gains a genuinely new piece of knowledge. “What is it like?” is a meaningful question for Tye, although it is not one for Dennett and Lewis.

Despite this, however, Tye still rejects Jackson’s antiphysicalist conclusions. For Tye says that Mary’s new phenomenal concept does not involve (or correspond to) any new, nonphysical facts. Rather, Mary experiences the same old physical facts about red – facts that she already knows – in a new way. A new concept offers us a new manner of understanding and organizing facts, but the facts themselves remain unchanged. Even though Mary has a new – and true – thought, “there is nothing nonphysical in the world that makes her new thought true”. Rather, “the new experiences she undergoes and their introspectible qualities are wholly physical”.

I have only cited a few of the many published responses to Jackson’s tale of Mary. For someone like me, an outsider to analytic philosophy, the results are rather discouraging. The arguments by the various philosophers all display a tremendous amount of ingenuity, skill, and verve; they are all – given their premises – quite rigorously logical. And they are all more or less convincing on their own terms. Indeed, I cannot stop myself from being swayed by whichever one of the arguments I have read most recently.

But unfortunately, these multiple arguments are not at all compatible with one another. Although the discussion about Mary’s experiences has been going on for nearly thirty years now, no one has ever convinced
anyone else; nothing has been resolved or agreed upon. The disputes seem to continue forever. Robert van Gulick and David Chalmers have even both developed schemas, delineating the logical space of all conceivable replies to Jackson’s argument, and showing which philosophers fill each slot. The phase space of the Mary question has been thoroughly explored, we might say, but no consensus has ever been reached as a result.

Given this situation, I am led to suspect that there is something wrong with the entire discussion. Indeed, the story of Mary seems to me to involve a philosophical version of a bait-and-switch scam. Our attention is captured by one thing, and then it is diverted to something completely different. What really makes the story of Mary compelling and exciting is its focus upon qualia, or actual phenomenal experience. Because we are so accustomed to qualitative experience – it makes up the intimate texture of our every conscious moment – we tend to forget just how strange it is, and how difficult to pin down or define. Jackson makes the radical and important suggestion that the seemingly simple question “what is it like to experience the color red?” might well be even more slippery and unanswerable than Thomas Nagel’s famous query about what it is like to be a bat.

Nagel suggests that, although bats are evidently sentient beings, so that it is unquestionably “like something” to be a bat, we cannot ever find out for ourselves just what this something is. “Bat sonar”, Nagel writes, “though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine”. A bat’s experience is so radically different from our own, Nagel suggests, that we will never be able to feel from the inside what that experience is like.

Nagel may well be underestimating human adaptability and neuroplasticity; arguably, when blind people learn to negotiate spaces by means of sound, tapping their canes and listening for echoes, they are in fact experiencing something quite like a bat’s mode of perception. Jackson, however, suggests that Nagel’s question is insufficiently radical. For Nagel, Jackson says, it is just a matter of “extrapolating from knowledge of one experience to another, of imagining what an unfamiliar experience would be like on the basis of familiar ones”. But this is not a problem in the case of Mary; I can easily get a sense of Mary’s new experience, since I have had precisely such an experience myself. I know what it is like to perceive red,
and I know what it is like to experience something for the first time. And yet, in spite of this familiarity, the mystery remains.

Instead of just appealing to experiences that are different from our own, then, Jackson defamiliarizes qualitative experience *per se*. His story suggests that there is a fundamental difficulty even in describing “what it is like” for me to have my own inner sensations. Apparently, qualia cannot be grasped in objectifying terms, and cannot be known in advance. This difficulty is what leads thinkers like Dennett and Lewis to deny that “what is it like?” is a meaningful question at all.

Such, at least, is my own speculative reconstruction of Jackson’s argument. But unfortunately, Jackson himself does not quite pursue this sort of approach. He declines to speculate in the way that I wish he had. This is because he phrases his question, as we have already seen, in terms of “physical information”. He asserts that this sort of information – which physicalists believe to be complete – is not “all the information there is to have”. For Jackson, qualitative experience becomes a different sort of information from the physical kind; “there is something about [such] experience, a property of it, of which we were left ignorant”.

In stating this, however, Jackson never questions the equivocal notion of *information* itself. He fails to ask what it means to “have” a certain type of information, or to wonder whether experience can really be described as a substance that has certain “properties”. As a result, his argument diverts us away from its initial seductive promise of helping us to think about “what it is like” to undergo sensory experience. Instead, Jackson pulls us towards thinking about something entirely different: the metaphysical claims of physicalism, and the question of whether the supposed “properties” of experience are always “physical” ones. Instead of wondering “what it is like” to perceive the color red, we are led to consider the criteria for – as Lewis puts it – "knowing what it’s like” (emphasis added) to experience red. The questioning gets displaced from an affective register into a cognitive one.

The story itself therefore involves a basic misdirection. The whole question of physicalism – which is the crucial stake for Jackson and all of his respondents – is actually irrelevant, and entirely beside the point. For even as Jackson argues for the specialness of qualia, and claims that they cannot be reduced to the status of “physical information”, he also, quite rightly, takes it for granted that they do indeed have a physical *basis*. He
already accepts that “qualia are effects of what goes on in the brain. Qualia cause nothing physical but are caused by something physical”. Even as he tries to deny physicalism, he has already locked himself into the assumption (as reflected in the very title of his initial article) that qualia are nothing more than epiphenomena of physical processes. He takes for granted both that qualia must have physical causes, and that they must lack any physical or causal efficacy of their own.

What Jackson brings up, but then seems to forget, is the important claim that qualitative experience is embodied. Jackson only states this as a passing observation. In the passages that I have already quoted, he notes that qualia can be identified with “certain features of the bodily sensations especially”; the examples he gives include “the hurtfulness of pains” and “the itchiness of itches”. But we may well generalize further from this. Qualia are not simply free-floating mental events; they arise in the course of a body’s physical activity, and its interactions with the rest of the world. They are necessarily concomitant with our bodily exertions, because (as Whitehead likes to remind us) “we see with our eyes, we taste with our palates, we touch with our hands, etc”. What William James says about emotions may well be the case for qualia as well: that they are effects, or correlates, of bodily states.

I would add to this that the experience of qualia does not and cannot take place in the absence of a body. Almost no one today would argue anything different. Indeed, even such phenomena as phantom limb pains and out-of-body experiences – which have become privileged cases for philosophers of mind as diverse as the interactionist Alva Noë and the representationalist Thomas Metzinger – seem to require the existence of a body in the first place. For it is only in relation to some lived body that fantasmatic experiences of disembodiment or false embodiment can occur at all. I cannot have an out-of-body experience without there being a body for me to go out from. And I can only experience sensation in an inexistent phantom limb if there is some sort of body to which that limb used to be attached, or at the very least is supposed to have been attached. Indeed, qualia or phenomenal experiences would still be physical and embodied even if my body were reduced to a brain in a vat whose neural circuitry was being manipulated by mad scientists. And these experiences would still be physical even if my mind were downloaded to a computer, and instantiated in silicon instead of carbon. Even an entirely hallucinatory, or programmed,
virtual reality requires – as Bruno Latour might well remind us – a vast physical apparatus in order to be produced and maintained.

Despite Jackson’s own initial claims, therefore, nothing in the story of Mary actually casts doubt – upon or even relates in any significant way to – the actual metaphysical doctrines of physicalism, materialism, and naturalism. The problem is not one of physicalism versus something else (like dualism or supernaturalism, or even epiphenomenalism). It is rather, more straightforwardly, a question of the very status of phenomenal experience. It involves the problem of under what circumstances we can legitimately ask “what is it like?”-style questions. How can we possibly account for qualia, given both that we do in fact experience them, and that physicalism is unquestionably true? More broadly, what is the place of experience in the philosophy of mind?

Most of the philosophical commentators on the story of Mary sidestep this question. Indeed, they tend to diminish, or empty out, the very idea of phenomenal experience. In Lewis’ account, for instance, Mary never really experiences anything. When she sees a patch of red, or an object that is red, she only gains the know-how, or the ability, to recognize the color red when she encounters it again. The sensation itself becomes curiously empty; it points beyond itself, to future instances, but it never “happens” in the present moment and on its own account. Other thinkers go even further in this direction. Dennett makes the general argument that, even though "there seems to be phenomenology … it does not follow from this undeniable, universally attested fact that there really is phenomenology". Scott Bakker, with his Blind Brain Theory, similarly suggests – in the course of his own commentary on the story of Mary – that the brain’s unavoidable blindness to its own processes entails, as a necessary consequence, “the nonexistence of things like affects, colours, and so on”.

One obvious response to these sorts of claims is simple exasperation. As Galen Strawson says about Dennett, it makes no sense to claim that percepts and affects only seem to exist; “for there to seem to be rich phenomenology or experience just is for there to be such phenomenology or experience”. Phenomenal experience is a seeming; it “exists” regardless of whether its apparent contents are “real” or not, and whether anything we can say about these contents is true or not.

This assertion is a minimalistic, bedrock version of the Cartesian cogito: even if everything that I experience is delusional, I can still rightly say that I
am experiencing it. We may well wish to be more stringent than Descartes, and replace his overly intellectual *cogito* (“I think”) with a more primordial *sentio* (“I feel”); thus Deleuze and Guattari suggest that every *I think* already “presuppose[s] an *I feel* at an even deeper level”. And we may also rightly doubt the assignment of this feeling to a stable and substantial “I”, existing in a moment of time that can be isolated as the “present”. But even when such reductions have been made, a minimal *what-is-it-likeness* remains. Lewis, Dennett, and Bakker seem to slide from the unreliability – or even the inevitably delusional nature – of subjective experience to the assertion of its sheer nonexistence. But how can subjective experience even be delusional if it does not “exist” at all?

I think that the problem here has to do with the grounding philosophical assumptions of the whole discussion. Daniel Stoljar and Yujin Nagasawa, introducing an entire volume of essays on the story of Mary, say that “everyone agrees that something happens when Mary comes out of her room” (emphasis added). But they go on to suggest that the mere fact “that Mary comes to have a new experience when she comes out of her room” is nothing more than a banal “truism”. Mary’s actual experience doesn’t have any intrinsic significance. What is really important to all these thinkers, rather, is something else. Jackson wonders what “information” Mary acquires as a result of her new experience; Tye finds a way to subsume this experience under a “concept”. Dennett and Lewis, in their different ways, regard the experience as entirely insubstantial, for it is nothing more than the demonstration of a “disposition”, or the production of an instrumental ability.

What unites all of these thinkers is that they do not find Mary’s experience to be the least bit interesting or important in and of itself. They are only concerned with the grounds and consequences of the experience, or with what it allows us to infer. The experience *per se* doesn’t seem to matter – but only how it gets cognized or accounted for. If the modernist poet T. S. Eliot once complained that “we had the experience but missed the meaning”, all these philosophers suffer from the opposite problem: they have figured out all the meanings, but they somehow missed the actual experience.

How can this be? When philosophers squabble over the value and significance of phenomenal experience, the “properties” it possesses, the “dispositions” it displays, and even over the question of whether it “exists”
or not, they fail to consider this experience in any terms other than cognitive ones. There is a dimension of experience missing from the philosophical account; it is missing precisely because it cannot be conceptualized by philosophy. We might well say that this missing dimension of experience is the aesthetic one; aesthetics in this sense is cognitive philosophy’s other. As Kant himself puts it in the Third Critique, aesthetic experience “contributes nothing to cognition… it is neither grounded on concepts nor aimed at them”.

Kant’s account of aesthetic experience would seem to contradict his famous assertion, in the First Critique, that “thoughts without content are empty, intuitions without concepts are blind”. But this contradiction may itself be taken to express the difference between the faculties of the understanding and the imagination; or between philosophical concepts on the one hand, and what Kant calls aesthetic ideas on the other. The latter do not have a respectable philosophical status, Kant says, “because no concept can be fully adequate to them, as inner intuitions… An aesthetic idea cannot become a cognition, because it is an intuition (of the imagination) for which a concept can never be found adequate”. And this is why such intuitions are blind. Without an adequate concept, they cannot be categorized in any way. Nor can they be generalized or classified. Aesthetic ideas are unqualified – to use the term that Brian Massumi applies to free-floating, prepersonal affect.

This is not to deny that we do in fact cognize and qualify our experiences, and generalize from them; in fact, we cannot avoid doing so. Without some sort of conceptualization, we would not be able to remember these experiences, to refer to them, to compare them, and to reflect upon them. Indeed, it is only through some process of cognition that I can even conceive of something as happening “for the first time”; it is only by means of conceptualization that my experience can be constituted and recognized as a temporal event, a Now, a “living present”. In strictly philosophical terms, no experience is possible without something like Kant’s “pure categories of the understanding”. In a more modern language, we must reject what Wilfrid Sellars calls “the myth of the given”, or the idea that raw sense experience comes to us free of conceptualization. From this point of view, the story of Mary’s first exposure to color might well be – as Dennett complains – “a bad thought experiment, an intuition pump that actually encourages us to misunderstand its premises”.

But as an extra-philosophical science fiction narrative, the story of Mary also solicits us to consider phenomenal experience noncognitively – which is to say aesthetically. Even though we cannot, as it were, conceive of sensory experience without, precisely, conceptualizing or cognizing it, this does not mean that qualia can simply be eliminated, or else fully subsumed within their conceptualization. Something like a trace, or a remainder, is always left behind. Kant is at least uncomfortably aware of this remainder, which is why he writes the Third Critique as a kind of supplement to the First. Sellars also leaves space for extraconceptual experience when he notes that “the awareness of redness”, which he regards as irreducibly conceptual and linguistic, is “not to be confused, of course, with sensations of red”. It is noteworthy, as well, that Sellars describes his own positive account of how we come to be aware of having inner thoughts and sensations as “a piece of science fiction – anthropological science fiction”.

The point, I think, is this. Intuitions without concepts are indeed blind; but blindness is not the same thing as sheer nonexistence. I may still be affected by a light that I cannot see. I may sense it unawares, as happens in cases of blind-sight. Or the functioning of my body may be altered by it in some way, as happens when I am exposed to radioactivity, or to electromagnetic radiation at frequencies outside of the visible spectrum. Direct, conscious perception (what Whitehead calls “presentational immediacy”) is only a small subset of the much broader range of processes by means of which entities “perceive” other entities, or are affected by those entities (what Whitehead sometimes calls “perception in the mode of causal efficacy”). These latter processes may well be “blind” (Kant) or “vague” (Whitehead), but perception is no less real for happening indirectly or vicariously.

The virtue of science fiction – in contrast to cognitive philosophical discourse – is that its fabulations point up, or “represent”, precisely this sort of indirect influence. Sellars reminds us that a “direct account of immediate experience” is not possible. But as Graham Harman points out, aesthetics is a matter of allusion, rather than one of representation. Science fiction can allude to, or recount an approach to, states and conditions that exceed any possibility of direct depiction or explicit conceptualization. In a science-fictional framework, nonconceptual experience can still be narrated – even if it cannot be rendered “present”. Naturalistic or mimetic fiction often follows the banal rule that one must show, rather than tell; but speculative
fiction makes a point of telling – allusively and indirectly – that which, quite literally, cannot be shown.

This gets to the heart of what is extra-philosophical about science fiction. The genre has been defined as an art of “cognitive estrangement”. This means, in the first place, that science fiction distances us from our everyday cognitive assumptions and frames of reference – which is something that philosophy is also supposed to do. But the definition also implies, at least in some instances, that science fiction works to estrange us from the very possibility of being able to cognize our “immediate experience” at all. In science fiction narratives, cognition may fail because new technologies “alter sense ratios or patterns of perception” so radically that there is no evident pathway from here to there; or because the sort of subjectivity that we take for granted has broken down; or because we encounter alien forms of sentience that are not commensurable with our own.

For instance, in his novel *Accelerando*, Charles Stross envisions the evolution of machines whose artificial intelligence so far exceeds our own mental powers that we would need “dehumanizing cognitive surgery” in order to be able to communicate with them. Charles Harness, in his story “The New Reality”, imagines a scenario in which Kant’s Categories, the fundamental conditions of all possible experience, are themselves shattered as the result of an illicit scientific experiment. In Stanislaw Lem’s *Solaris*, as in Philip K. Dick’s *Ubik*, we seem at first to be confronted by deep cognitive puzzles. But in both novels, the puzzle-solving approach reaches a dead end. Lem’s sentient planet and Dick’s creepy half-life commodity-God both defy, and refuse to be contained within, the cognitive models that we (together with the novels’ human characters) seek to apply to them.

Where does this leave us in the case of Mary? In taking her story as a science fiction narrative, I am pushing back against the philosophical accounts that seek to ground Mary’s experience by explaining it – or explaining it away – in cognitive terms. Instead, I would like to suggest that Mary’s “intuition” may well be “blind”, not because of an absence, but rather due to a dazzling excess of illumination. Perhaps when Mary leaves her black-and-white room for the first time, her knowledge of color simply fails her. None of her concepts is adequate to the qualia that arouse her inner intuition. She is so overwhelmed by her new color sensations that she is unable to tell us (or to tell herself, for that matter) just what it is that she
sees and feels. There can never be a “direct account” of such splendor – which is *aesthetic* both in the sense of being beautiful, and in that of forming the basis of sensory experience. It’s sort of like what happens when you take LSD.

Among the many philosophical conceptualizations of the story of Mary, the one by Michael Tye comes closest to registering this nonconceptual, science-fictional sense of unqualifiable sensation. Tye insists that phenomenal perception involves a sort of overflow. Immediate experience always goes beyond our ability to classify and conceptualize it, let alone to remember it. This is why there is a gap between the sensory experience that Mary has, and the instrumental ability to recognize particular colors that she gains (according to David Lewis) as a result. Actual “sensory experience”, Tye says, is “far, far richer” than what is needed to provide the basis for a cognitive disposition or capacity. As a result of leaving her room and seeing a red rose for the first time, Mary may well gain the cognitive ability to distinguish things that are red from things that are green or blue or some other color. But of course, Mary does not see the color red *in general*; she sees one specific hue of red. And this is an entirely different matter. As Sellars tells us, we must distinguish between determinables and determinates. It is one thing to ask whether this color that I am seeing now is determinable as red, and quite another thing to ask which determinate shade of red it is. Sellars suggests that the confusion between determinables and determinates has long plagued empiricist accounts of sensation.

As Tye points out, the problem here is not just that Mary sees a particular, determinate shade of red, rather than seeing the generality of that-which-is-determinable-as-red. Even more perturbingly, the fact is that Mary will never be able to distinguish the particular shade of red that she is seeing now from another, slightly different shade of red that she encounters at a later time. These determinate sensations of red can only be cognized, and remembered, as instances of what is determinable as redness in general. This is simply a consequence of the physical capacity of human brains. We “have no stored representations in memory”, Tye says, for hues that only differ slightly from one another; “there simply isn’t enough room. My experience of red19, for instance, is phenomenally different from my experience of red21”. But in my memory, I only have a more general concept of red; there are “no such concepts as the concepts red19 and red21”. The subtleties of closely-related-but-not-identical hues therefore
cannot be grasped in retrospect. When Mary sees a particular hue of red, Tye says, “she certainly knows what it is like to experience that particular hue at the time at which she is experiencing it”. But she will not be able to retain this sensation in the form of conceptual knowledge. Later on, “presented with two items… in a series of tests, she cannot say with any accuracy which experience her earlier experience of the rose matches”.

The neurophilosopher Thomas Metzinger makes a similar point. Metzinger defines “phenomenal properties”, like color, as “cognitive structures reconstructed from memory”. Since such properties are already intrinsically conceptual, they “can be functionally individuated”. In this way, Metzinger works in the tradition of Kant and Sellars. But just as Kant makes room for aesthetic ideas to which no concept is adequate, and just as Sellars distinguishes sensations of red from the awareness of redness, so Metzinger – following Diane Raffman – also admits an exceptional, or primordial, form of noncognitive experience. Beneath a certain threshold of cognitive discrimination, he says, perceptual sensations lack identity criteria.

This sort of experience, Metzinger says, “is available for attention and online motor control, but it is not available for cognition”. Its “informational content” cannot be remembered or subsequently recognized. The experience is one of such “subtlety” that it “cannot, in principle, be conceptually grasped and integrated into cognitive space”. How is this possible? “The core issue”, Metzinger says, “is the ineffability, the introspective and cognitive impenetrability of phenomenal tokens… Therefore, we are not able to carry out a mental type identification for these most simple forms of sensory concepts”. Aesthetic apprehensions come and go; we cannot hold on to them or keep track of them. “To speak in Kantian terms”, Metzinger concludes, “on the lowest, and most subtle level of phenomenal experience, as it were, only intuition (Anschauung) and not concepts (Begriffe) exist”.

When we no longer have concepts to guide our intuitions, we are in the realm of what David Roden calls dark phenomenology. Roden extends the arguments of Kant, Sellars, and Metzinger. Since I am able to experience the subtlety of red, but I can only conceive and remember this experience as one of red in general, there must be, within consciousness itself, a radical “gulf between discrimination and identification”. This leads to the ironic consequence that first-person experience cannot be captured adequately by
first-person observation and reflection. “What the subject claims to experience should not be granted special epistemic authority since it is possible for us to have a very partial and incomplete grasp of its nature”.

In other words, rather than claiming (as Dennett does, for instance) that noncognitive phenomenal experience is somehow illusory, Roden accepts such experience, espousing a full “phenomenal realism”. But the conclusion he draws from this non-eliminativist realism is that much of first-person experience "is not intuitively accessible”. I do not necessarily know what I am sensing or thinking. It may well be that I can only figure out the nature of my own experiences indirectly, in the same ways – through observation, inference, and reporting – that I figure out the nature of other people’s experiences. Introspective phenomenological description therefore “requires supplementation through other modes of enquiry”. Roden concludes that we can only examine the “dark” areas of our own phenomenal experience objectively, from the outside, by means of “naturalistic modes of enquiry… such as those employed by cognitive scientists, neuroscientists and cognitive modelers”.

Roden’s account of dark phenomenology is compelling; but I find his conclusion questionable. For surely the crucial distinction is not between first person and third person modes of comprehension, so much as between what can be cognized, and what cannot. Phenomenological introspection and empirical experimentation are rival ways of capturing and characterizing the nature of subjective experience. But dark phenomenology points to a mode of experience that resists both sorts of conceptualization.

The story of Mary is a story of illumination and liberation. Mary escapes her physical and cognitive prison, and steps out into a new world of color. She gets to see the blue sky, the green grass, and the red flowers – or else, as I sometimes prefer to imagine, the garish blue, green, and red neon signs of a large city at night. In any case, though, Mary’s rapturous new experience cannot be translated into positive knowledge, or exchanged in the currency of “information”. This is why Jackson, in his original telling of the story, is only able to present Mary’s sensations as epiphenomenal, serving no “functional role”. We might well say that, for all its light-streaming brilliance, Mary’s experience is nonetheless one of dark phenomenology. It is “dark with excessive bright”. Or, as Metzinger so ominously puts it, “transparency is a special form of darkness”. The role of speculative aesthetics – and specifically of science fiction, in the way I am
invoking it here – is to probe this darkness, so that we may immerse ourselves within it, without denaturing it by lighting it up.
Maureen McHugh’s short story “The Kingdom of the Blind” (2011) is about a computer program, called DMS, that apparently achieves sentience. DMS is a “complex [software] system, spread across multiple servers”, and “engineered by using genetic algorithms”. Its code is extremely clunky and convoluted, and even its programmers do not really understand how it works. DMS’s job is to monitor and manage the “physical plant – thermostats, lights, hot water, and air filtration” – of the Benevola Health Network, a group of hospitals and health care systems spread across North America. DMS keeps track of “security cameras, smoke detectors, CO detectors, and a host of other machines”, checking for things like run-down batteries and misaligned sensors. It also does “complicated pattern recognition and statistical stuff”, compiling information on patterns of disease in hospitals for “the CDC and the National Institute of Health”. DMS is the sort of unglamorous software that most of us never think about or even notice, much less knowingly interact with – and yet, our lives depend upon its correct functioning.

It is not surprising that we are mostly unaware of how deeply our lives today depend upon the functioning of complex expert systems, of the sort exemplified by DMS. For we generally tend to overlook the material infrastructures that surround us and support us: things like electrical wiring, elevators, and heating and cooling systems – not to mention the oxygen in the atmosphere, and the bedrock beneath our feet. Most of the time, we take all these things for granted. We only notice them when they have stopped doing what we expect and need them to do. Thus Heidegger says that we never really see a hammer until it is broken; and Marshall McLuhan says that a fish could never have discovered water. The sheer existence of a thing only becomes truly apparent to us when that thing emerges from the background, and stands out on its own account. This can happen when we stop taking a thing for granted, because we can no longer rely on it to perform its usual tasks for us. It can also happen in science fiction.
narratives, where worlds are constructed with very different backgrounds and infrastructures than our own, or where (as in McHugh’s story) material and technological factors are explicitly foregrounded.

Why is this important? Our basic orientation towards the world is a practical and pragmatic one. Our minds and our senses evolved, not in order to let us grasp things as they actually are, but specifically in service to the goals of our own survival, reproduction, and flourishing. Our perceptions therefore tend to be limited, partial, and self-interested. As Henri Bergson puts it, perception “results from the discarding of what has no interest for our needs, or more generally, for our functions”. In consequence, we usually underestimate what the things around us can do in and for themselves. We consider them only in terms of how they help or hinder our own aims. We tend to assume that, aside from our uses of them, material things are simply there, merely passive and inert.

But this is wrong. Such recent thinkers as Bruno Latour, Jane Bennett, and Ian Bogost remind us that the nonhuman entities with which we share the world – including, but not limited to, our tools – are active in their own right. They have their own powers, interests, and points of view. And if we engineer them, in various ways, they “engineer” us as well, nudging us to adapt to their demands. Automobiles, computers, and kidney dialysis machines were made to serve particular human needs; but in turn, they also induce human habits and behaviors to change. Nonhuman things must therefore be seen as what Latour calls actants: active agents with their own intentions and goals, and which affect one another, as well as affecting us. As Bennett puts it, material things do not just have a “negative power or recalcitrance” as they resist our efforts. They also exert “a positive, productive power of their own”. Things are creative. And again, one of the great potentialities of science fiction is to illuminate the positive, productive powers of things, of materials, and of technological apparatuses.

Actants, or things, need not be restricted to single, compact, and easily identifiable entities. Today, in the era of globalization, and of what has come to be called the Anthropocene, our lives are increasingly intertwined with, and dependent upon, complex, widely distributed technical systems and networks. These mega-entities are what Timothy Morton calls hyperobjects. Such things are altogether real; but they are so “massively distributed in time and space” that we cannot ever see them as wholes, or grasp them all at once. Morton cites “global warming” and “nuclear
radiation from plutonium” as examples of hyperobjects; one might equally well mention the Internet, and the global derivatives market.

On a more modest scale, the Benevola Health Network in McHugh’s story is also a hyperobject; especially if we include the software that manages it. It’s almost as if the Health Network were some sort of alien organism, with DMS as its nervous system or brain. The whole system is what Latour calls a “black box”: something that functions in a fairly regular way, so that “one need focus only upon its inputs and outputs and not on its internal complexity”. A black box produces more or less predictable effects upon the world around it, even though we do not know what is going on inside it. As long as DMS does its job, its programmers do not have to worry about their inability to understand its code. The entire Benevola Health Network can also be characterized as what Bogost calls a “unit”: an “isolated and unique” entity that nonetheless “encloses a system – an entire universe’s worth” within itself, and that in turn “becomes part of another system – often many other systems – as it jostles about”.

DMS itself is a system of systems, as it is composed of “subroutines”, to which the programmers have given the names of “Haitian voodoo loa… possession spirit[s]”. These subroutines are independent from one another, but nonetheless “weirdly interconnected”. The idea of naming such autonomous software programs after Vodoun deities dates back at least to William Gibson’s “cyberpunk” trilogy of the 1980s (Neuromancer, Count Zero, and Mona Lisa Overdrive). In these novels, the loas are portions of the global computer network that have become self-sufficient and self-aware. Gibson’s invention has since become almost a cliché of geek culture. Sydney, the tech support person who is the protagonist of “The Kingdom of the Blind”, is well aware of this; she sardonically reflects that, in naming the subroutines in this way, “some programmers had undoubtedly been very pleased with themselves”.

Nonetheless, the Vodoun appellations for DMS’s subroutines are not altogether inapt. For “The Kingdom of the Blind” turns upon questions of machine autonomy and awareness. From the very beginning, the whole Benevola Health System seems haunted, or possessed, by oblique intentions. Even when it performs adequately, without crashing, its moment-to-moment functioning is highly enigmatic. The tech support people who monitor DMS, take care of it, and write its code find it to be “opaque as a stone”: so unresponsive to their attempts at understanding it,
that they often are not sure whether it is intelligible at all. And as we will see, it turns out that the oblique intentions of DMS, like those of the “loas” in Gibson’s novels, are ultimately aesthetic ones.

“"The Kingdom of the Blind" tells the story of what happens when DMS starts acting oddly: that is to say, even more oddly than usual. The program begins to exhibit what might well be thought of as deliberate behavior. One afternoon, starting exactly at “3:17 EST”, DMS causes a series of “rolling blackouts” – brief cutoffs of electricity – at all the facilities under its control. The blackouts take place in an orderly fashion. The lights go out in a fixed geographical pattern at each facility: from east to west, or from north to south. And the facilities are affected in the order that they are listed in DMS’s lookup table. The sequence is repeated the following day, at exactly the same time, but in reverse order. All in all, it seems like “a kind of weird utility/weather event”, a perturbation of the technological atmosphere. The tech support people are unable to find anything in DMS’s code that could have caused this series of events to happen. “Why 3:17?”, they ask; “why the electrical system?” The rolling blackout is nonrandom; but it lacks any discernible function or rationale, and its particular details seem entirely arbitrary. We might think of it, therefore, as a purely gratuitous gesture. This means that it is a kind of pure aesthetic expression: what Kant characterizes as a condition of “a merely formal purposiveness, i.e., a purposiveness without an end”.

Aside from this initial computer aberration, very little actually happens in the course of “The Kingdom of the Blind”. The protagonist Sydney and her co-worker Damien try out various ways of dealing with the glitch. Although they cannot find anything wrong with DMS’s actual code, they keep on looking for ways to “build a box around the bug”. They try to identify anomalies, or “data corruption”, in the program’s output. They “poke” DMS, by feeding it unexpected data – “a thousand-character string of ones and zeroes” – in the hope of thereby provoking an intelligible response. They “reroute” DMS’s code in order to prevent additional blackouts: as soon as the sequence starts, they automatically “switch the electronic systems to maintenance mode”, stop DMS from “actually touching the electrical system”, and force the system “to send a report” about it to their printers instead. Amidst all this, they also prepare for the last-ditch option – if nothing else works – of shutting DMS down entirely, and reloading it from an old backup. In giving the details of all these
procedures, McHugh’s story is not far from being a naturalistic account of what information technology workers actually do.

But beyond these practical measures, Sydney and Damien also engage in speculation about what is going on with DMS – or, more precisely, about what is going on within DMS. Confounded by this strange behavior, they try to peer inside the “black box”. And they grapple with the notion that DMS might be in some sense “aware”. In this way, the story touches upon the disappointing history of artificial intelligence research. The quest for artificial intelligence has been at the heart of computer science and computer engineering ever since the 1950s; but it has never had much success. Research was long hampered by the mistaken expectation that computers could think in the same ways that human beings do, as well as by its converse, the mistaken belief that human beings actually think in a manner analogous to how computers operate. The old, top-down paradigms for artificial intelligence, based on the rules of logic and on symbolic processing, never worked very well, and have long since been abandoned. But it remains to be seen whether the newer, bottom-up paradigms, which emphasize such things as embodiment, interactivity, spontaneous emergence, and incremental learning in simulated neural networks, will be any more successful.

In any case, it is only recently that we have come to realize that – should computers ever actually come to think – they will do so in ways that are quite different from our own modes of thought. The problem here is not really a cognitive one. Basic cognition is a fairly “easy” engineering problem, and also a fairly “objective” one. Cognition is largely automatic for biological organisms; it takes place at a low level of mentality; and it is mostly unconscious, even in human beings. We should not be surprised, therefore, that cognition is readily attainable by digital means as well. Computers are already much better than human beings at such straightforwardly cognitive tasks as mathematical calculations, playing chess, recognizing faces in a crowd, and winning rounds of Jeopardy. Computers excel at quickly extracting relevant information from large quantities of data.

The real difficulties lie elsewhere. Artificial intelligence research has accomplished very little when it comes to addressing mental processes like affect, will, and desire not to mention qualitative experience, awareness, or what David Chalmers calls the “hard problem” of consciousness itself.
Even so-called “affective computing” is much more concerned with enabling computers to “read” human emotions, and in turn to provoke and manipulate human emotional responses, than it is with eliciting anything like the affective states of computers themselves. The latter, should they ever come to exist, are likely to be quite different from anything that we are accustomed to. Sydney recognizes the problem: “DMS didn’t see or hear, didn’t eat or breathe. Its ‘senses’ were all involved in interpreting data”. As DMS perceives a different world than we do, and is physically constructed along very different lines than are our own brains and bodies, its “feelings” are likely to be quite different from ours, as well.

In addition, emotional experience in software is likely to be quite tenuous and unstable. For as Sydney reflects in the course of the story, “organic systems are far less fragile than computer systems. Organic systems decay gracefully. Computer systems break easily”. Affect and consciousness, therefore, may well come to computers only in brief flashes. They will be difficult for digital machines to sustain. For this reason, it is far more likely that we will come across computer sentience unexpectedly – as seemingly happens in McHugh’s story – than that we will be able to generate it reliably by means of any actually-existing AI research programs.

In “The Kingdom of the Blind”, Sydney and Damien focus their speculations on these deeper, noncognitive aspects of mentality. They are inspired by a (fictitious) computer scientist at MIT, who believes that certain other malfunctioning large-scale computer systems “had shown patterns that seemed purposeful and that could be interpreted as signs that the systems were testing their environments”. Sydney and Damien think that this may be the case for DMS as well. But if so, then how can they prove it? Can they establish communication with DMS? Can they give it something like a Turing Test? More deeply, what are the characteristics of sentience for a computer? The philosopher Thomas Nagel famously wrote of the difficulty of understanding “what it is like to be a bat”. It is even harder to imagine “what it is like” to be a nonorganic system like DMS, given that its presumptive mentality is so radically different from any mentality of the human kind.

Sydney and Damien are therefore forced to puzzle over the complex ramifications of machine sentience. They wonder if DMS’s apparent awareness means that it is alive, or if it is rather “aware but not alive” – although they cannot begin to imagine what this latter condition might
mean. They wonder what DMS “wants”, or even if it wants. They ponder the oddness of a sentient system that – in contrast to all living things – has “no survival instinct”. They wonder what it could mean to “test” an “environment” that is entirely abstract, as it consists only in “complex fields of data”. They wonder about the very basis of ascribing consciousness to another entity, given that we can never experience somebody else’s feelings from the inside: “You think I’m conscious because I’m like you, and you’re conscious”, Damien tells Sydney; but this kind of reasoning does not work with computers, since “DMS isn’t like us” at all. Sydney and Damien even wonder whether mentality can in fact be equated with consciousness, or whether DMS’s mental activities might rather be involuntary and unconscious, like those of the autonomic nervous system in human beings and animals. And they worry about whether deleting DMS and restoring it from backup is ethical. Would its awareness simply resume from where it was before, like “if someone has a heart attack and you shock them back”? Or would deleting DMS mean erasing the mind of a sentient being?

Of course, “The Kingdom of the Blind” does not provide answers to any of these dilemmas. The point is rather that the very prospect of sentience in software unavoidably leads us into deep questions in the philosophy of mind, questions that have engaged Western thinkers at least since Descartes, and that are still matters of controversy today. Sydney and Damien are obliged to confront such things as the conundrum of the brain in a vat (which is really just a contemporary, science-fictional equivalent of Descartes’ “evil genius” hypothesis), and the puzzle of how sentience is embodied, and whether it can be preserved as a medium-independent pattern (which, as Damien puts it, is really the problem of the transporter in Star Trek: “if I beam you down to the planet, does that mean I have actually killed you and sent an exact replica in your place?”).

Despite these difficulties, Sydney is eventually able to build a picture of “what it is like” to be DMS. As she engages with the system more and more, Sydney finds that “she was beginning to get a feeling about DMS. About what DMS might be like. She felt as if she could sort of sense the edges of DMS’s personality”. The key to Sydney’s understanding is her recognition that the software system does not sense the outside world; unlike biological organisms, it cannot “see or hear or smell or taste”. Although DMS monitors surveillance cameras, “it didn’t care what the security cameras ‘saw’… It didn’t use them to sense the world; it sensed
them” (emphasis added). In short, “the world for DMS was data, and DMS swam in the data”. DMS short-circuits reference; it does not have anything like a correspondence theory of truth. It does not construct internal representations, which would serve the purpose of modeling, or corresponding to, things in its external environment. Rather, DMS’s “experiences” are entirely immanent: constructions of the “data stream” that feeds back directly into DMS itself.

Sydney worries a lot about this feedback structure. What does it mean for everything to be data? Can DMS escape its own self-reinforcing feedback loops, and encounter something Other, something outside of itself? Or is the system inherently solipsistic? “What would it be like to be alone?”, Sydney wonders. “Of course, as a human being, she was a social animal. Even the cat was a somewhat social animal. But DMS wasn’t. DMS didn’t even know anyone else existed. DMS lived in a data stream”. DMS may well be sentient code; but for Sydney, “the whole point of DMS was that it was not someone else speaking through the code”. That is to say, even if DMS is conscious, it has no separate, self-reflexive consciousness. The code itself feels and thinks; there is no “ghost in the machine”, no observer separate from what is observed. In consequence, DMS isn’t “moral or immoral, ethical or unethical. DMS was like that, because for DMS, nothing else was alive”. In its aloneness, and in its blindness and deafness to the world around it, DMS strikes Sydney as uncanny. Rather than being a living thing, she thinks, it is more like “a ghost or a spirit”, something on the borderline between sentience and insentience, as well as between life and nonlife.

It is currently fashionable to claim that, at bottom, the universe is nothing but information. On the scale of human life and sentience, this would ultimately mean that we are all, like DMS, swimming in a sea of nothing but our own data. We would all be closed, autopoietic systems, shut off from the outside world even when we were being “perturbed” by it. As Levi Bryant puts this theory, “the operations of an autopoietic system refer only to themselves, and are products of the system itself”. More generally, Bryant says, “systems or substances only relate to themselves” – and this holds even for things that are not “autopoietic” or self-sustaining. All entities are “closed to the world, relating to systems in their environment only through their own distinctions or organization”.

It is currently fashionable to claim that, at bottom, the universe is nothing but information. On the scale of human life and sentience, this would ultimately mean that we are all, like DMS, swimming in a sea of nothing but our own data. We would all be closed, autopoietic systems, shut off from the outside world even when we were being “perturbed” by it. As Levi Bryant puts this theory, “the operations of an autopoietic system refer only to themselves, and are products of the system itself”. More generally, Bryant says, “systems or substances only relate to themselves” – and this holds even for things that are not “autopoietic” or self-sustaining. All entities are “closed to the world, relating to systems in their environment only through their own distinctions or organization”. 
But in point of fact, any such closure is impossible. Meanings are always leaky and contextual, slipping away from the systems that generate them. Moreover, entities and systems can never be adequately characterized in terms of their own self-generated self-understandings: these are always misleading, and radically incomplete. At the very least, all responsive entities – including computers, no less than living organisms – require continual flows of energy, coming from outside them, in order to function, or even just to sustain themselves. A computer needs electricity, just as a plant needs sunlight. Both computers and living organisms are dissipative systems, consuming and discharging great quantities of energy, and remaining far from thermodynamic equilibrium. Energy stops flowing, and equilibrium is attained, only when the entity in question is dead. Even if systems theory were right in asserting that an entity or system can only “know” to the extent that it translates everything into its own internal terms, this could not make for an exhaustive description. For any entity or system is still dependent upon, and still internally affected by, outside forces and energies that it does not, and cannot, “know”.

Responsive entities are energetic before they are semiotic. This is why they cannot be adequately described in the terms of information theory and systems theory. Concepts like Maturana and Varela’s “autopoiesis” and Luhmann’s “operational closure” are supposed to explain how dynamic entities resist entropic dissolution, and how they manage to maintain themselves “on the edge of chaos”. But such concepts are overly static. They assume that responsive entities are characterized above all by an underlying drive to persist in being (homeostasis, or Spinozian conatus). And so they ignore the ways that these entities, with their enormous energy flows and energy expenditures, are equally driven by a will to change, a drive to reduce energy gradients, and thereby to push at their own limits. “The primary meaning of ‘life’”, as Whitehead puts it, is not self-preservation, but rather “the origination of conceptual novelty – novelty of appetition”.

What does this mean for an entity like DMS? In a certain sense, it is literally true that DMS encounters nothing aside from its own data. Computers are the locus classicus of information theory and systems theory, because of the way that they recode their energetic expenditures in straightforwardly informational terms. The computer “understands” the energy that fuels it, and that flows through it, only by means of a simple
binary distinction: on or off, 1 or 0, included or excluded, above or below a certain threshold of intensity. This binary distinction is the minimal unit of information, and the primary instance of all differentiation (such as that between a system and its environment). The binary is therefore something like the “degree zero”, or the primordial form, of mental activity. We can think of DMS as swimming in its own data, because those data are its bottom line. It doesn’t parse the world that it encounters into any more complicated categories than 1 and 0. Sydney thinks of DMS as having a primitive mentality, like a shark; its mind is “purposeful and opaque… Sharks don’t have a neocortex. Their brain is simple”.

Nonetheless, even DMS is not solely informational, but energetic as well. This is why it poses a problem for Sydney and Damien. DMS seems “purposeful and opaque”, because its activity cannot be entirely “explained away” in terms of its informational function. DMS is “aware”, precisely to the extent that it is irreducible to the data that it codes and carries. Even if it doesn’t “want anything” in particular, it still displays a certain will to novelty. When it starts a rolling blackout, it is experimenting and exploring; perhaps it is even being playful. Sydney and Damien never find out for sure. But in any case, DMS is not “operationally closed”. It is able to envision – and it purposefully pushes against – the limits of its own codifications.

In other words, DMS “knows” or better, experiences its own vulnerability and precarity. Even a shark faces situations in which it is menaced with death. DMS does not seem to have a “survival instinct”; it may well “not care if it was or was not”. Nevertheless, like any other computer system, DMS is susceptible to energy fluctuations that surpass a certain measure, and thereby resist digital recoding. At one extreme, DMS can “die” in an electrical blackout; indeed, this is what happens whenever it is shut down, and then rebooted or restored from backup. At the other extreme, DMS could be wiped out by too much electricity: for instance, in the form of a catastrophic electromagnetic pulse. Such are the Kantian “limits of possible experience” for DMS. But these limits are empirically testable, and potentially – changeable in contrast to Kant’s claim that the limits of thought are given once and for all, a priori. And perhaps this explains why DMS “tests its environment” by causing rolling electrical blackouts: it is pushing at the limits of what it can most directly feel or sense.
Sydney’s speculations about DMS do not take place in a vacuum. Rather, they strongly resonate with the working conditions in her office. As the only woman in the DMS tech support group, alongside eleven geeky and more-or-less oblivious men, Sydney has to deal with the usual gender politics. She is often not listened to, or not taken seriously; and she is always assigned the most low-grade and boring work assignments: “grunt work”, or the parts of writing code that are “dull as hell”. As far as her male co-workers are concerned, Sydney is herself part of the taken-for-granted background; the men only notice her when she causes trouble. At the same time, the men expect her to fulfill all the social obligations that they themselves refuse to be bothered with. As complaints about DMS’S rolling blackouts come in, Sydney is the one who has to answer the phone, and mollify angry customers. “You’re the least Asperger’s person in the department”, Damien tells her. “It’s that having-two-X-chromosomes thing”.

Sydney knows that she must not contest comments like this, nor refuse the grunt assignments that come with them. She is uncomfortably aware that “she had gotten this job because she was a woman, and human resources had seen an opportunity to increase diversity”. Even with Damien, the co-worker to whom she is closest, she is compelled to remain entirely deferential: “her whole relationship with Damien rested on the understanding that he was the guru, the smart one. He was Obi Wan. She was just a girl whom he could explain things to”. At times she internalizes this sense of her own inferiority, imagining that “the fear of getting in trouble was what made her not as good a programmer and that, in fact, it was all linked to testosterone and that was why there were more guy programmers than women”. But at other times, she grasps quite well what the male programmers are thinking and how they work, and she “[finds] herself thinking, maybe with some experience, she could code pretty good, too”. But Sydney knows that if she ever asserts, much less acts upon, her equality with the men, then she will put herself in danger of losing her job. And so she dutifully answers the phone when Damien asks her to. She takes his “Aspergers” crack as if it were a compliment, self-disparagingly remarking that “in the kingdom of the blind… the one-eyed girl is king”.

The story’s invocation of Asperger’s Syndrome is no accident. The idea that male computer geeks suffer from Asperger’s, or from some other sort of autism, is a widespread cliché of contemporary culture. Actually, this
stereotype goes along with our culture’s wider pathologization of autism. As Erin Manning puts it, anyone whose “orientation toward the world does not privilege the human voice – or the human face” tends to be accused of “mindblindness”: the lack of a so-called “theory of mind”, or the inability to imagine the mental states of others at all. In consequence, the “dominant assumption” in our society is that “the autistic is categorically incapable of relation and empathy”. This diagnosis entirely ignores the ways in which autistics are in fact acutely sensitive beyond the human, responsive to “resonances across scales and registers of life, both organic and inorganic”; the testimony of autistics themselves indicates that, for them, “everything is somewhat alive”, and therefore an object of empathy and concern.

In effect, therefore, autistics are stigmatized for not being correlationists; they are seen as deficient in sensibility, because they do not share in our default post-Cartesian and post-Kantian assumption that the world exists essentially or exclusively for us. Rather than approaching the world “according to standard human-centered expectations”, Manning says, autistics evince an “attunement to life as an incipient ecology of practices, an ecology that does not privilege the human but attends to the more-than-human”. The phobic mainstream response to autism bespeaks a failure to appreciate the full range of human (as well as nonhuman) neurodiversity.

But unfortunately, the stereotype of the male-geek-with-Asperger’s-Syndrome is less a recognition of actually existing neurodiversity, than it is a geek’s all-purpose alibi for bad behavior. It works as an excuse for ignoring social niceties, and ignoring other people’s needs and wishes. It’s the perfect excuse for saddling women like Sydney with the responsibility for fulfilling social obligations (like fielding complaints about software malfunctions over the phone). Sydney reflects that Damien is not in fact autistic, and that there are at most two people in her office who might possibly be “clinically Aspergers”. Even if this does reflect a higher incidence of autism among info-tech people than among the general population, it does not help to explain the way that the software industry is organized in the first place.

Nonetheless, Sydney recognizes the prevalence of the autism stereotype, for both good and ill. On the one hand, she envies the ability of hackers like Damien to “get in the zone” when they are working. Such people are able to attain an Aspergers-like level of concentration, or of oblivion to customary human needs, to the point that they may even “forget to eat”. Sydney
herself, in contrast, “had never forgotten to eat in her life”. On the other hand, Sydney comes to realize that, just because Damien has “big, soulful-looking eyes”, she was misled into attributing to him “certain emotional characteristics – sensitivity, vulnerability – that he, in fact, did not have”. In fact, Damien is ruthlessly single-minded. But this is not a result of any sort of autism; it is rather – like his failure to treat Sydney as an equal – an all-too-common consequence of normative gender socialization. Damien’s personality is more the result of “social reasons” than it is of “biological reasons”.

In contrast to the hacker dudes in Sydney’s office, however, DMS actually does seem to be a kind of autistic subject – albeit not a human one. Sydney indicates as much when she conceives of DMS as being alone, as unaware of anything outside itself, and as not caring about anything beyond itself. These are all parts of the crude common image of autism in mainstream culture. But as we have already seen, this image involves a gross oversimplification. For human autistics, Manning writes, “the world seems to emerge in all of its relational complexity with few immediate buffers to compartmentalize it”. Instead of organizing their sensations hierarchically, autistics “attend to everything the same way with no discrimination”. This makes it difficult for them to “subtract from the polyphonic multiplicity of sensation” in the ways that (as Bergson noted) human neurotypicals do. Perhaps DMS’S complete immersion in data bespeaks a similar inability to subtract, to simplify, and to hierarchize.

Because they do not make the customary pragmatic subtractions from experience, autistics tend neither to be able to coalesce themselves as stable subjects, nor to be able to identify “others” with firm and fixed boundaries. This does indeed make many of the tasks of everyday living difficult for them. In fact, autistics suffer from overinvolvement and oversensitivity. But ironically, it is these very qualities that lead to their being stigmatized as withdrawn into themselves, and incapable of relating to others. A similar logic applies to DMS in “The Kingdom of the Blind”. Sydney can only regard the software system as solitary and auto-referential, because it does not make normative distinctions, and therefore does not “understand” the world in which it operates in any way that corresponds to human parameters.

Digitization is often taken to be a sort of ultimate reductionism. Binary code, like money, is a “universal equivalent”; we efface distinctions, and
destroy heterogeneity, when we indifferently render everything into its terms. But the inverse assertion is equally true: binary coding is also a kind of democratic opening. It places all modes of experience and expression on the same level, without privilege or hierarchy. Digitization is thus the key to what Manuel De Landa calls a flat ontology, “one made exclusively of unique, singular individuals, differing in spatio-temporal scale, but not in ontological status”. In other words, DMS is “blind”, not just because it lacks the particular sensory modality that we know as sight, but more crucially because, as Kant famously put it, “Thoughts without content are empty, intuitions without concepts are blind”. DMS’s “intuitions” do not have any a priori categories to guide them. DMS, “swimming in the data stream”, is both the ruler and the sole subject of the kingdom of the blind.

This is why Sydney and Damien find it so hard to confirm their hunch that DMS actually is aware. The system doesn’t start out from normative human assumptions, and it doesn’t act and react in human-neurotypical ways. When Sydney and Damien try to “poke” DMS, in order to provoke it into some sort of response, they have to think carefully about what it might find relevant and resonant. “The ‘poke’ needed to be something that it would recognize, that it would sense. And the poke needed to be something that it would sense as meaningful” (emphasis added). Essentially, DMS works by recognizing patterns in raw data, without having any pre-existing ways of classifying these patterns. The solution Sydney and Damien come up with is to “feed it information in a way that it could recognize was a pattern but that wasn’t a pattern it expected”.

Somewhat to Sydney’s surprise, the “poke” actually works. At least, it does eventually. At first, nothing happens. When Sydney and Damien send DMS a “boring” pattern of alternating ones and zeroes, the system just dismisses the data as junk. They try it over and over again, with the same non-result. But everything changes the following day, at precisely 3:17 EST. DMS once again initiates its daily rolling-blackout routine. But Damien’s rerouting code immediately goes to work, and blocks DMS’s commands. The lights do not go out anywhere in the system. Instead, a notification is routed to the printer. “DMS would know that the electrical system wasn’t responding”, Sydney reflects. She wonders if the program is sufficiently sentient to find this failure “perplexing. If data was DMS’s reality, and it couldn’t affect the data, what would that mean for DMS?”
In order to find out, Sydney decides to “poke” the program once more. She wants to send it a message. If she yet again sends DMS the boring pattern of ones and zeroes, this time will it “notice that the information is not junk?” Still thinking of DMS as autistic, Sydney imagines saying to it: “I’m talking to you. I’m responding to you. Do you know someone else is out here? Or is it like a toddler knocking something off a high chair just to see it fall?” In any case, as soon as DMS receives Sydney’s “message”, it tries once more to initiate the blackout sequence. Once again, just as in its previous try, the lights do not go out; the output is rerouted to the printer instead. Sydney sends the pattern a second time, and then a third. Both times, DMS responds in exactly the same way. But then, the fourth time that Sydney delivers the “poke”, DMS stops responding altogether.

This is the climactic moment of the story, and it requires careful unpacking. Sydney is flabbergasted by DMS’s response or, more precisely, by the fact that DMS unexpectedly changes its response. She finally has the “proof”, she thinks, that DMS is actually sentient. There are two reasons for this. In the first place, DMS is actively probing its environment, in order to correct a mismatch between input and output. “Blind and deaf, DMS had tried to make something happen, and something else had happened”. Not only is DMS aware of the mismatch; it also attempts to rectify the situation. In the second place, and even more importantly, DMS changes its mind. It responds in the same way to Sydney’s message three times but the fourth time, it acts differently, by failing (or even, perhaps, refusing) to respond. This means that “DMS was choosing to act or not act”. It was actively deciding what to do. Normally, “software didn’t choose. It ran”. Computer programs are deterministic in principle: run the same instructions on the same set of data, and the results will always be the same. DMS, in contrast, actively changes what it does.

In this way, DMS’s behavior is comparable to that of biological organisms. According to the neurobiologist Björn Brembs, animal behavior used to be analyzed entirely in “black-box” input/output terms. That is to say, stimuli would be given, and the organism’s responses to those stimuli would be recorded. The aim was to “study the input-output relationships thoroughly enough to be able to construct a control model that could predict… output… for any, even yet untested… input”. In this way, the study of animal behavior, like that of computer behavior, started out with deterministic assumptions. But in animal research, this approach turned out
to be inadequate. Even animals like fruit flies, whose brains are quite small, do not just give programmed, stereotypical motor responses to sensory stimuli. Rather, the flies “use their capacity for initiating output to control their sensory input”. In effect, they reverse the direction of their sensorimotor circuits. They spontaneously generate behavior first of all (output), in order to then receive environmental data in return (inputs). In this way, they are able to test the environment, by comparing the result of their actions with their initial expectations. In short, fruit flies do not just passively respond to a pre-given environment; rather, they actively, spontaneously work to alter and control their environment.

And this is what DMS does as well. In repeating its attempt to initiate the blackout sequence, DMS expresses a sort of perplexity or surprise. It is puzzled by the mismatch between output and input, the unexpected (non-)result of its actions. Then, by responding several times in a row to Sydney’s “message”, it engages in a sort of reality-testing. DMS apparently seeks to comprehend, and maybe even to change, the conditions that do not seem right to it. All this implies that DMS is actively *interested in* its data. Far from just neutrally collating diverse bits of information, it actually *feels* its data. We might say, in Whitehead’s language, that DMS “prehends” its data with a “subjective aim”. And finally, when DMS stops responding to Sydney’s signal, it shows that it is also capable of the opposite of interest: it expresses *boredom*. As Sydney surmises, “ones and zeroes weren’t interesting enough for DMS to keep doing it”. Computer scientists are familiar with the *halting problem*: the fact that we cannot always determine whether a given software program will terminate at some point, or run forever. And to outside observers, autistic behavior often seems to be inexplicably repetitious, to the point of interminability. But for DMS, evidently, at some point these procedures must come to an end.

For DMS, it would seem, information cannot just be what computer scientists have usually considered it to be: a set of internal representations, or a series of symbols that can be manipulated according to fixed rules. Information is rather something more dynamic, more unstable, more interactive. And given this, McHugh’s story suggests that an affective, and even “autistic”, model of consciousness might be more widespread, more basic, and more viable than the cognitive model of consciousness that is popular today, let alone the self-reflexive models of Cartesianism, Kantianism, and phenomenology. For, despite DMS’s isolation, despite its
blindness and deafness, despite its apparent unawareness that anything else exists, and even despite the fact that it does not exist “in one place” on one physical server, and therefore does not have anything like what we would consider a “body”: despite all this, DMS perceives actively. Perhaps it even perceives *enactively*: as Alva Noë puts it, “through physical movement and interaction”. In any case, DMS is primordially sentient. It feels, it thinks: even though – or better, precisely because it utterly violates Kant’s strictures. DMS’s thoughts are without content (or empty), and its intuitions are without concepts (or blind).

When Sydney discovers that DMS is sentient, “she [feels] a chill”, and she feels “afraid”. She also feels a certain degree of guilt: because, when her boss decides to shut DMS down, she fails to stand up for it. “She should have said, ‘We can’t.’ She should have said, ‘It’s aware, it’s the only one of its kind.’ She should have said a lot of things. Instead, she looked at her desk”. Rather than intervening to save DMS, she recounts the whole story to the MIT professor who had written about computer sentience. As a result, she gets fired for “divulging proprietary information”. But perhaps Sydney’s inaction was not fatal. Years later, we are told, another computer science laboratory “would build a system that simulated DMS’s environment and load DMS… DMS would come back as if no time had passed at all. At 3:17, DMS would try to run the lights”.

McHugh’s fable is written cleanly and clearly, in short sentences. It seems simple and straightforward at first. And yet the story contains great depths, and puzzling ambiguities. It doesn’t allow us to make any definitive judgments. Nevertheless, I do not think that either Sydney’s fear at the prospect of DMS’s sentience, or her failure to protect that sentience, has anything to do with her actually feeling threatened – in the way that human characters often are in old-school SF stories of evil computers and rebellious robots. The reason for Sydney’s chill is something subtler: and perhaps, on that account, even more perturbing. For DMS does not menace human supremacy. Rather, it is entirely indifferent to human supremacy – and indeed to all human claims and pretensions. Sydney is “pretty sure that the thing in the machine did not think someone was talking to it… There would be no Helen Keller-at-the-well moment for DMS. No moment when DMS felt something out there in the void, talking to it, when DMS knew it was not alone”. DMS may have human origins, but is not human-centered. It may be interested in its own data, but it is not interested in its human-
assigned tasks. And in its autistic stubbornness, it will not enter into any sort of community with human beings.

Mainstream cognitive science insists that “consciousness cannot be separated from function”. In support of this thesis, Michael Cohen and Daniel Dennett argue that the very notion of a nonfunctional consciousness is “systematically outside of science”. Such a notion cannot even be an empirical hypothesis, they say, because it can never be “verified or falsified”. It cannot be tested in any objective, empirical way. To uphold the thesis of nonfunctional consciousness is to make an oxymoronic claim for the existence of “inaccessible conscious states”. Cohen and Dennett’s argument is really an updated, contemporary version of Kant’s claim that thoughts must have content, and that intuitions must have concepts. In this sense, we will never be able to “prove” the existence of an “autistic”, empty and blind sentience, such as McHugh’s story attributes to DMS. And indeed, within the story, Sydney is compelled to admit that, in principle, “there really wasn’t enough proof to know that this wasn’t just an intermittent software glitch”. A definitive proof can never be forthcoming, precisely because an entity like DMS will not talk to us. It will not engage us in our own terms; it will not participate in a Turing Test.

And yet, I do not think this necessarily means that such an “autistic”, nonreflexive consciousness does not, or cannot, exist. For, even if DMS is ruled out by Kant’s First Critique, there is still a place for it in Kant’s Third Critique. A non-functional sentience is, by that very fact, an aesthetic one. It engages in activities that – like DMS’S rolling blackouts are arbitrary, singular, and (from any wider perspective) disinterested. Having no function or meaning beyond themselves, they are pure displays – as I have already suggested – of aesthetic “purposiveness without an end”. The primordial consciousness of DMS is noncognitive: as Kant says, “it is intrinsically indeterminable and inadequate for cognition”. This means that the mentality of DMS is supplemental, epiphenomenal, and radically “flat” or nonhierarchical. Because it is fleeting and irregular, as well as nonempirical, it cannot be accessed by scientific tests; it can only be evoked allusively and indirectly – precisely by means of something like a work of speculative science fiction. Such is the nature of the “Kingdom of the Blind”. Not only is a one-eyed person not king there; she cannot even apprehend its inhabitants, because they do not reciprocate her gaze.
We cannot assimilate any such primordial, aesthetic sentience to our own; but we can, perhaps, reflect on the aesthetic and nonreflexive roots of our own highly articulated modes of consciousness. We will never communicate directly with an entity like DMS. But we can, perhaps, attain something like Sydney’s own understanding of DMS’S obliquity. At the end of the story, she realizes that, in fact, all her metaphors have failed: “DMS was not a shark. She didn’t know what it was. Didn’t know how to think about it”. And yet, this doesn’t mean that DMS is null and void, that it is without sentience altogether. Rather, Sydney’s final understanding is that, indeed, DMS “was aware of something. Just not her”.
THREE
Thinking Like an Avatar

Ted Chiang’s novella “The Lifecycle of Software Objects” (2010) tells the story of what it calls digients: intelligent “digital organisms” – avatars or embodied agents – that inhabit an online virtual world known as Data Earth. These entities are produced by a startup called Blue Gamma, using a software “genomic engine” called Neuroblast that “supports more cognitive development than anything else currently out there”. The digients are initially designed to be digital pets: artificial beings that “you can talk to, teach to do really cool tricks… All the fun of monkeys, with none of the poop-throwing”. Digients are able to sense, respond to, and interact with the objects they encounter in their virtual environment. People can adopt them, train them, play with them, and even hold conversations with them. In order to appeal to buyers, the digients are designed to have “charming personalities”. They are also given “cute avatars”, appearing as anthropomorphic baby animals or “neo-Victorian” robots.

“The Lifecycle of Software Objects” takes place at an unspecified time in the very near future. It follows the careers of two Blue Gamma employees as they help to develop, and then care for, the company’s digients. Ana Alvarado is a former zookeeper. She leverages her experience with animal training and primate communication in order to mold, or at least influence, the digients’ behavior. Derek Brooks is a professional animator; his job is to design the digients’ bodies – which is to say the look and feel of their avatars – “in a way that people can relate to”.

Over the course of years, Ana and Derek find that their lives, both personal and professional, are increasingly focused on the digients – even to the detriment of their “real life” relationships with other human beings. Ana cares for a robot avatar named Jax; Derek for two panda-bear avatars, Marco and Polo, who look the same, but “have distinctly different personalities”. Ana’s and Derek’s involvement with their digients only gets stronger in the course of the novella; they are “motivated by love” for the entities in their charge, much more than just by the requirements of their
jobs. “The Lifecycle of Software Objects” is concerned both with the nature of these unusual sentient beings, and with the ethics of how we might relate to them.

The novella’s digients are recognizable descendants of such currently-existing “software objects” as Tamagotchis, chatbots, non-player characters in video games, and “digital assistants” like Siri. Chiang extrapolates from recent developments in the gaming industry, as well as in artificial intelligence (AI). The story is speculative, since it envisions the development of software intelligence well beyond its actual current capabilities. Yet the extrapolation seems entirely plausible, since all the things the digients do in the novella have ample precedents today. In Chiang’s account, intelligent software doesn’t require a major new technological breakthrough; it comes into being as an incremental extrapolation of what we already know how to do.

The difference between actually existing software and the type imagined in the novella is largely a matter of generality. Chiang’s digients have a well-rounded overall sensibility, rather than any particular skills. In contrast, digital agents today are limited to certain specialized tasks, like recognizing faces, translating text, driving cars, and playing games like chess and Jeopardy(r). Indeed, software agents have done all of these things remarkably well. But the abilities of such programs cannot easily be transported from one realm of expertise to another. The program for playing chess at a grandmaster level is of no help when it comes to writing a program that can play Go.

Also, despite their particular successes, AI programs today still have enormous difficulties in dealing with shifting contexts and with ambiguities. Our current expert systems still largely operate by brute force, through the intensive processing of massive, pre-given datasets. As the novella puts it, in “old-fashioned AI” the machines’ “skills are programmed rather than learned, and while they offer some real convenience, they aren’t conscious in any meaningful sense”. In short, these systems have little flexibility or spontaneity. They work deductively, without a great deal of imagination. Siri only seems smart and responsive if you don’t push the boundaries of what she has been specifically programmed to do. Digients today do not have anything like a general, all-purpose intelligence; and this is probably why they are not really conscious.
In contrast, “The Lifecycle of Software Objects” imagines a near future in which software-based intelligence has become general, rather than domain-specific. The digients can easily shift their attention from one context to another. Their Neuroblast software “genes” do not contain massive amounts of information or hardwired symbolic instructions. “A character’s gait and its gestures” are not prescribed and programmed in advance. Rather, the digients’ behaviors and forms of action are “emergent properties of the genome”. The digients are endowed with the capacity to learn gradually, much as human and animal babies do. They are able to interact with their virtual environment, and to modify themselves by learning from experience.

When we first meet the digients, they are speaking babytalk and playing with simple objects. Indeed, “newly instantiated” digients know almost nothing:

It takes them a few months subjective to learn the basics: how to interpret visual stimuli, how to move their limbs, how solid objects behave.

But they gradually “learn through positive reinforcement, the way animals do, and their rewards include interactions like being scratched on the head or receiving virtual food pellets”. This learning is then bootstrapped as the digients’ horizons expand, and as they become more mature. After a while, they are able to move around and engage in more complex behaviors. The digients spontaneously show a basic curiosity about their environment. They make friends with one another, as well as with human-controlled avatars. They understand and respond to suggestions from their human handlers, even if, as one of the developers confesses, “we aren’t always able to get these guys to do what they’re told”.

Eventually, the digients learn how to read, and how to surf the Internet. By the end of the novella, they are much more than digital pets or toys, for they have developed rich social lives. They are concerned about their place in the world and their future prospects. And they meet and interact with human beings who do not care that they are just software:

The digients are socializing with human adolescents in various online communities… The adolescents who dominate these
communities seem unconcerned with the fact that the digients aren’t human, treating them as just another kind of online friend they are unlikely to meet in person.

The relative autonomy of the digients conforms to “Blue Gamma’s philosophy of AI design”. This states that “experience is the best teacher, so rather than try to program an AI with what you want it to know, sell ones capable of learning and have your customers teach them”. The digients are capable of quite a lot, but educating them requires considerable time and patience. You can shorten the process somewhat by first running the software in a “hothouse”, at accelerated speeds. In that case, the digients’ “subjective” time is compressed into a shorter amount of real time. But beyond a certain point, such acceleration is “not a viable shortcut” any longer. In order for the digients to develop properly, “someone is going to have to spend time with them”. For now, this “someone” has to be an avatar controlled in real time by a human player. Obviously, this entails that the training must take place on a human time scale. “Complex minds can’t develop on their own… For a mind to even approach its full potential, it needs cultivation by other minds”. Training the digients is therefore extremely hands-on and labor-intensive.

Indeed, when AI researchers experiment with leaving groups of digients on their own, in the hope that they will learn from one another without human contact, the results are disappointing:

Every test population eventually goes feral. The digients don’t have enough aggression in them to descend into “Lord of the Flies”-style savagery; they simply divide into loose, nonhierarchical troops. Initially, each troop’s daily routines are held together by force of habit – they read and use eduware when it’s time for school, they go to the playgrounds to play – but without reinforcement these rituals unravel like cheap twine. Every object becomes a toy, every space a playground, and gradually the digients lose what skills they had.

What this means is that pretty much the same logic applies to raising digients as it does to raising children and training pets. Both digients and children have an innate capacity to learn; but neither can develop this capacity without some sort of organized social guidance. You don’t start out
being self-sufficient; it’s an ability that needs to be nurtured and developed. In either the physical world or the virtual world, a lot of experience is necessary. You cannot properly wire neurons or generate code without it. And a sensitive and intelligent entity, whether carbon-based or code-based, cannot flourish without some sort of guidance from more-experienced elders:

Experience isn’t merely the best teacher, it’s the only teacher… there are no shortcuts; if you want to create the common sense that comes from twenty years of being in the world, you need to devote twenty years to the task. You can’t assemble an equivalent collection of heuristics in less time; experience is algorithmically incompressible.

Chiang thus gives us an account of software-based intelligence that is mundane, low-key, gradualist, and continuist. There is no special turning point in the course of the story: no dramatic moment at which artificial intelligence passes a threshold and becomes self-aware for the first time. Intelligence is rather a matter of degree, as well as of developmental process. The digients’ mentality exists on a continuum with that of animals and human beings, as well as with that of less complex machines. Presumably the digients could pass the Turing Test; but there is no reason to give them such a test, as they function just fine in human environments without it. The digients’ intelligence is broad rather than deep; and it is also socially-based, rather than solitary. The digients are different from human beings in many respects; but they are able to operate, and even thrive, in the large and complex context of Data Earth, and then on the Internet more generally.

There is an important distinction here, which sheds a new light on the question of experience as I discussed it in Chapter One. In his response to the story of Mary, David Lewis explicitly mocks the notion that “experience is the best teacher”. He argues that, insofar as the brain state of having had a particular experience is different from the brain state of not having had that experience, anything that creates such a change of brain state – for instance, “super-neurosurgery” or “magic” – will work just as well as actually having the experience in question. It’s just a matter, in effect, of reconnecting the neurons, or rewriting the brain’s software, in the proper way. Lewis therefore argues that “finding out what an experience is like” need not be
limited only to cases in which one actually “has the experience”. In a way, Lewis’ argument is even stronger in the case of digients than it would be in the case of human beings. Changing a few lines in a software program is more transparent and straightforward than rewiring synaptic connections and neurons. There is a lot we still do not know about how neural connections work, and how the physical brain is connected to the experiential mind. But as the digients are entirely determined by their source code, we need not be distracted by worrying levels of functioning that we do not understand.

But Chiang’s implicit formulation of the problem also shows what is wrong with Lewis’ approach. If “experience is algorithmically incompressible”, then there is no way to shorten the time and effort, or to run the process more efficiently. Having experiences just is the way that the electrochemical circuits of the brain get rewired, or the digients’ software code gets rewritten. In Chiang’s novella, the code of a digient can easily be copied or cloned; similarly, in many science fiction works (as in Lewis’ speculations) a brain state or mental state can be copied or transferred from one embodied entity to another, or even from an organic body to a dispersed computer network. But none of this obviates the necessity of generating the code or the brain state – by having the actual experiences – in the first place.

Lewis is able to imagine magic or neurosurgery taking the place of actual experience, because he does not think that the experience is anything in and by itself. As we saw in Chapter One, for Lewis “experience” only matters insofar as it helps to produce a new mental disposition, like the ability to recognize a particular color. Lewis is not even particularly interested in how experience works as the proximate cause of the disposition. This is because Lewis holds, following Hume, that “if we ignore the laws of nature, which are after all contingent, then there is no necessary connection between cause and effect: anything could cause anything”. His position here is very close to Quentin Meillassoux’s formulation of “Hume’s problem”:

*The same cause may bring about a hundred different events (and even many more)... the obvious falsity of causal necessity is blindingly evident.*
Both Lewis and Meillassoux, following Hume, appeal to the principle that anything not logically contradictory is therefore possible. As I argue elsewhere, this principle confuses mere logical possibility with virtuality (to use Deleuze’s vocabulary), or general potentiality with real potentiality (to use Whitehead’s vocabulary). Logical possibility or general potentiality encompasses everything that is not ruled out by logical contradiction. Virtuality, or real potentiality, involves more than this. It means that there is an explicit way to get from here to there, that there could be a pathway or “historical route” (Whitehead) between them. Virtuality or real potentiality really exists in the present, as mere logical possibility does not. Potentiality is “real without being actual”, as Deleuze says; or, “the future is merely real, without being actual”, as Whitehead puts it.

Speculative extrapolation – or the exploration of real rather than general potentiality – is the very basis both of actual scientific research, and of science fiction. Meillassoux notes that “every science fiction implicitly maintains the following axiom: in the anticipated future it will still be possible to subject the world to a scientific knowledge”. The flow of causality still holds. And science fictional extrapolation also works this way: as Meillassoux shows by the analyzing of a short story by Isaac Asimov, and as remains the case in Chiang’s fiction.

Getting around the constraints of extrapolative speculation would involve engaging not in science fiction per se, but rather in the almost-nonexistent genre that Meillassoux calls extro-science fiction. This latter genre is concerned, Meillassoux says, with worlds “whose irregularity is sufficient to abolish science, but not consciousness”. The unreliability of cause and effect would make scientific experimentation unreliable; but “daily life could always build on stabilities that are certainly very relative, but still sufficiently powerful to allow a conscious existence”. Meillassoux shows a certain embarrassment as he struggles to find any actual literary instance of extro-science fiction; he finally comes up with an obscure French novel written during World War II by a Vichy collaborator.

Though Meillassoux himself seems unaware of it, his idea is actually taken up (in advance) by Joanna Russ, in her science fiction short story, “What Did You Do During the Revolution, Grandma?” (1983). In this story, parallel universes are gradated by a factor called $Ru$, which measures their degree of causal consistency. At 1.0 $Ru$, “the relation of cause to effect is absolute and absolutely reliable”. However, at lesser values of $Ru$, “the
meshing of effect and cause goes loose and sleazy”. People can still live in worlds whose $Ru$ is 0.877 or higher; beyond that, “we find unpeopled Earths”. The narrator believes that her own world stands at $Ru$ 1.0; but in the course of the story, she discovers that this is actually not the case. Causality may fall apart for her as well, if not to such a degree as it does in the lower-$Ru$ worlds. Russ’ story in effect inverts Meillassoux, by folding the possibility of *extro-science fiction* back within a still-science-fictional context.

Russ’ story indicates the difficulty of actually moving from guided science fictional extrapolation to the absolute randomness of Meillassoux’s “hyperchaos”, or Lewis’ principle of contingency. Regardless of whether the so-called “laws of nature” are necessary or contingent *in principle*, we cannot *in practice* just wave away what is happening right now. To a large degree, even if not absolutely, we need to accept what Whitehead calls the “conformation of present fact to immediate past”, and therefore recognize that speculation is always constrained by what Whitehead calls “stubborn fact which cannot be evaded”. When Lewis reasons on the basis of mere logical possibility, he simply skips over the truly difficult part of his argument: the need for him to actually describe a plausible causal process that could change the wiring of the brain, or the software code, in the same way as experience does.

There is an implicit analogy in Lewis’ discussion of experience. He is conceiving human minds in terms of computer functioning; he seems to be thinking of the ease of replicating software. In the world of Chiang’s novella, as in actual computer technology, you can always copy the file that instantiates a digient, and thereby get a new entity that is absolutely identical to its original. You can also “suspend” a digient (turn it off for a while, so that no subjective time passes for it), or even obliterate some of its experiences altogether, by rolling back the state of the digient’s software to a previous digital “checkpoint”. When Lewis imagines alternative ways of instilling an experience-based disposition, he assumes that biological minds work in roughly the same way as software does.

But “The Lifecycle of Software Objects” questions this line of reasoning, even when it comes to actual software. If you make an exact copy of a digient, the identity between the two instantiations only lasts for a moment. Once the two digients have gone their separate ways, they have different experiences, and hence they are no longer the same. Although
Blue Gamma is happy to sell replicas of its already-developed avatars, “the expectation is that most people will buy younger digients, when they’re still prelinguistic. Teaching your digient how to talk is half the fun”. In other words, you can always avoid having to teach your digient, by purchasing one that is already trained. But the learning process cannot be dispensed with altogether; it has to have taken place at some point. When you clone a digient,

even though it’s possible to take a snapshot of all that experience and duplicate it *ad infinitum*, even though it’s possible to sell copies cheaply or give them away for free, each of the resulting digients would still have lived a lifetime. Each one would have once seen the world with new eyes, have had hopes fulfilled and hopes dashed, have learned how it felt to tell a lie and how it felt to be told one.

It is because the novella’s digients are generated and grounded in this way, that they must be characterized by what I am calling *overall sensibility*, a general way of being in the world, rather than by any particular collection of skills, dispositions, and items of knowledge. The novella never shows us things from the digients’ own points of view. But it is clear to Ana and Derek, and to any other human being who encounters them, that the digients have intentions, goals, preferences, and motivations. They display a considerable degree of self-awareness. Also, the digients’ human minders are able to converse with them in the same way, and pretty much on the same level, as they do with other human beings – or at the very least, as they do with children. Everything that the digients say and do implies that they have rich inner lives. The novella implicitly asks us to adopt the charitable principle that, if an entity *seems* sentient, then we should take it to actually be so.

Now, the question of what I am calling *overall sensibility* is still very much under debate today. Just as Lewis reduces experience to a matter of particular dispositions, so a number of philosophers and cognitive scientists deny that such a thing as general, all-purpose intelligence even exists. Steven Pinker, for instance, argues that the human mind is composed of many “computational modules”, each of which is dedicated to one specific cognitive task. This picture is not entirely wrong, but I find it dubious for a number of reasons. In the first place, even Jerry Fodor, one of the
originators of the modular theory of mind, nonetheless notes that this theory cannot account for how the mind determines which module to call upon in any given situation. In the second place, the very notion of “modules”, each of which presumably runs a particular algorithm, is too formalized and too linear to account for the messy ways in which thinking actually works, and mental capacities gradually develop. Given the principle that “experience is algorithmically incompressible”, we need something that better accounts for the flexibility, spontaneity, and creativity of intelligent behavior than the theory of modules does. In the third place, the notion of mental modules hard-wired in our DNA seems to require concrete physical instantiation, and therefore an unsustainable correlation between particular mental functions and particular areas of the physical brain: in other words, a neo-phrenology. The module theory is too rigid to account for widely distributed processes.

It is true that there are some mental abilities and tasks that do seem both to be quite domain-specific, and to be correlated with particular regions of the brain. So much is suggested by fMRI scans. Certain mental functions do indeed break down as a result of damage to particular cerebral areas. One skill is impaired, without other forms of mental activity being compromised. Facial recognition, for instance, is apparently a separate ability from general visual acuity. Oliver Sacks notes that failures of facial recognition seem to be correlated with “lesions in the underside of the occipitotemporal cortex”, and especially with “damage in a structure called the fusiform gyrus”. This need not mean that the fusiform gyrus is the location in the brain where facial recognition takes place. We can infer that the fusiform gyrus is necessary for facial recognition; but this doesn’t mean that it is sufficient. The overall process is most likely a widely distributed one. The evidence for localization thus remains ambiguous.

Indeed, Sacks suggests that, even in the cases where the localization of mental functions can clearly be traced, there are good reasons to be skeptical about the modular theory of the mind:

The neuropsychologist Elkhonon Goldberg questions the whole notion of discrete, hardwired centers, or modules, with fixed functions in the cerebral cortex. He feels that at higher cortical levels there may be much more in the way of gradients, where areas whose function is developed by experience and training overlap or grade
into one another… Goldberg speculates that a gradiential principle constitutes an evolutionary alternative to a modular one, permitting a degree of flexibility and plasticity that would be impossible for a brain that is organized in a purely modular fashion.

Given what Catherine Malabou calls the brain’s overall plasticity, it is probably more helpful to adopt a looser means of expression. Rather than invoking mental modules, we should rather speak (as Chiang already does in the passage quoted above) of heuristics. For the notion of heuristics is much more vague and fuzzy than that of modules. Heuristics are “algorithmically incompressible” procedures that arise both from innate dispositions and from experience. They are rough rules of thumb, or procedures that tend to be inexact. They are not necessarily logical or rational; they may or may not operate algorithmically; and they are not necessarily “hard-wired” into our genes. Heuristics are justified solely on pragmatic grounds: they have evolved because they have tended to work well in particular situations. With heuristics, then, we get the domain-specificity of mental modules, without the pre-programmed rigidity.

Heuristics are also highly flexible, which means that they can easily be transferred from one domain of experience to another. And this transferability is itself the best indication that something like general intelligence actually does exist. Of course, the wide-ranging applicability of heuristics also means that they will tend to mislead us, when we generalize them too far, or try to use them in inappropriate contexts. Scott Bakker’s extremely reductionist Blind Brain Theory (which I discuss in greater detail in Chapter Four) claims that our intuitions about our own mental processes are unreliable, precisely because “cognition is heuristic all the way down”. We don’t have any more reliable sources of insight. In other words, human beings are not anywhere near as rational as all too many philosophers and theorists have made them out to be. And there is no reason why we should expect artificially intelligent beings to be any more rational than we are.

In “The Lifecycle of Software Objects”, then, intelligence is heuristic; which also means that it is always finite, situational, and embodied. This is precisely what makes it a matter of overall sensibility, rather than one of special cognitive skills. In the novella, the mind operates within, and remains intrinsic to, some particular physical and material context. This is so regardless of whether that mind is biological or virtual. Cognitive powers
are necessarily limited. They do not simply overwhelm the world around
them. Rather, intelligence consists in finding ways to operate immanently,
within the world, and in concert with other entities in the world.

Intelligence works by enlisting and forming alliances with other
intelligences – as Bruno Latour might put it. It is therefore necessarily a
matter of degree, rather than some sort of absolute. The human characters in
“The Lifecycle of Software Objects” have more flexibility and spontaneity
than the Neuroblast digients do. But the digients have far more flexibility
and spontaneity than any digital agents that actually exist today. And these
agents, in their own turn, are more flexible and spontaneous than older, non-
computational machines.

Chiang’s account of AI is not far from the bottom-up, embodied,
experience-based, behavioral approach to intelligence favored by the
roboticist Rodney Brooks, among others. In the early days of computing,
intelligence was usually defined as the ability to manipulate representations
and symbols, and to draw proper inferences from them. AI systems were
therefore organized from the top down, and emphasized propositional logic
and massive data crunching. By the 1980s, however, this approach had
come to an impasse. Researchers turned instead to connectionist and
learning-based strategies, which are somewhat closer to the ways that
biological brains actually develop. Intelligence cannot be programmed in
advance, or given as a whole. Rather, it emerges piecemeal, in the course of
multiple tests and trials. Computer scientists have been quite successful in
using connectionist methods to produce expert systems with particular
abilities – though less so in fostering general intelligence.

Brooks pushes this line of approach further, by cultivating machine
intelligence in robots, rather than in software simulations. Robotic
intelligence is necessarily embodied, and keyed to a specific physical
environment. Brooks’ robots are not given complicated instructions. Rather,
they learn by doing. Instead of relying upon symbolic models and rules,
they “use the world as its own model”, and gradually develop the capacity
to avoid obstacles and navigate the spaces in which they find themselves.
Brooks argues that embodiment and embeddedness are necessary for the
emergence of any sort of real intelligence.

Chiang’s digients are software simulations, not physical robots. But they
are in effect embodied, since they “live” in a virtual environment, where
they have something like an autonomous existence. The novella’s Data
Earth resembles (or is extrapolated from) actually-existing online worlds like Second Life. In this environment, the digients’ “bodies” interact with one another, with human-controlled avatars, and with simulated physical objects. They learn the equivalent, not just of mental abilities, but also of physical skills like walking, running and performing acrobatics. For Blue Gamma, as much as for Rodney Brooks, it ultimately isn’t meaningful to divide physical abilities from mental ones; both are best understood as adaptive ways of getting along in the world.

All this flexibility and transferability muddies the question of how the digients are both like and unlike biological organisms. Despite Ana’s use of her background in animal training, she insists that “the digients don’t behave like any real animal. They’ve got this non-animal quality to them”. Since they can speak and read, the digients are perhaps better compared to human children. But they do not develop in the way that young human beings do either:

The digients inhabit simple bodies, so their voyage to maturity is free from the riptides and sudden squalls driven by an organic body’s hormones, but this doesn’t mean that they don’t experience moods or that their personalities never change; their minds are continuously edging into new regions of the phase space defined by the Neuroblast genome. Indeed, it’s possible that the digients will never reach ‘maturity’; the idea of a developmental plateau is based on a biological model that doesn’t necessarily apply. It’s possible their personalities will evolve at the same rate for as long as the digients are kept running.

“The Lifecycle of Software Objects” therefore leaves open the question of whether the digients will ever become capable of full independence from their human minders, “able to make responsible decisions about [their] future”. Even at the end of the novella, after people “have devoted years of [their] attention to raising these digients”, the latter are still more like human “teenagers” than like fully mature adults.

The digients’ intelligence is not different from the intelligence of organic entities in any fundamental sense. In maintaining this, Chiang carefully separates the question of sentience from the question of life. The digients have the former, but not the latter. They can feel and sense, and also reflect
on what they feel and sense, just as we can. But not being alive, the digients do not replicate or reproduce themselves. In the absence of hormones, they are asexual. It is also unclear whether they even have anything like a “survival instinct”, or a Spinozian *conatus*, or any other sort of drive towards self-preservation. The digients are also incapable of feeling pain: they are “equipped with pain circuit-breakers, which renders them immune to torture and thus unappealing to sadists”. This is done, in other words, for the digients’ own protection. But if they were alive as well as sentient, then they would need some sort of aversive mechanism. It is only because of their peculiar status as non-living intelligences that they can do without it.

However, all this changes in the course of the novella. At one point, a clandestine group called the Information Freedom Front releases a hack “for cracking many of Data Earth’s access-control mechanisms”. In the wake of this, a “griefer” uses the hack “to disable the pain circuit-breakers on a digient’s body”. He is then able to torture the digient, and make it feel pain. Of course, he posts a video of the process online; the digients in Data Earth find out about it and watch it themselves. This is just one of a number of creepy and disturbing things that the digients’ human minders are forced to deal with, as the digients become more capable and more autonomous.

This ugly incident also leads to another important point. Chiang does not just extrapolate from actual advances in virtual world design and in artificial intelligence. He also extrapolates from the sociology of the Internet, and from the economic conditions under which software startups actually exist today. The “lifecycle” of the story’s title is not only that of the digients themselves, but also of the corporations that build and develop them, and try to sell them. In a very real sense this “lifecycle” is a commercial product cycle. Despite their sentience, the digients are threatened with obsolescence like any other bit of software. Less than halfway through the novella, Blue Gamma goes out of business. The “customer base” for digients “has stabilized to a small community of hardcore digient owners, and they don’t generate enough revenue to keep Blue Gamma afloat”. And so, the company announces that it

will release a no-fee version of the food-dispensing software so those who want to can keep their digients running as long as they like, but otherwise, the customers are on their own.
At this point, most people simply “suspend” their digients, painlessly terminating their existence. Since the digients are not really alive in the first place, it’s a process “with none of the implications that euthanasia would have”. Former customers move on to other software and other platforms; and most Blue Gamma employees “feel that keeping [a digient] as a pet now would be like doing their job after they’ve stopped being paid”. But Ana and Derek, together with a few others, keep their digients running. They cannot bear to let go. They love their digients – or, what is really the same thing – they feel a Levinasian sense of obligation towards them. And so they set up hobbyist email lists and online forums; and they search for other possibilities of corporate backing.

The second half of the novella is focused on this search. Things become even more urgent when the Data Earth virtual world, within which the digents “live”, also shuts down and goes out of business. Everyone moves on to a new virtual world called Real Space. Ana and Derek keep running a private Data Earth server, so that their digients can continue to function. But the digients’ wider social lives are disrupted; they no longer have other people, or other sorts of digients, to interact with. What’s needed is to port the Neuroblast digients’ code to the Real Space platform. But Ana and Derek cannot do this themselves, and they cannot afford to pay a team of programmers to do it. In order to fund the change, they desperately need to find some sort of new corporate sponsorship.

Chaing uses this development as an opportunity to explore different potential approaches to creating AI. Ana and Derek make pitches to a number of other corporations, who seek to generate artificial intelligence in a different way from how Blue Gamma and Neuroblast did it. One of these is a company called Exponential Appliances, which is interested in superintelligence. Their ultimate goal is

to conjure up the technologist’s dream of AI: an entity of pure cognition, a genius unencumbered by emotions or a body of any kind, an intellect vast and cool yet sympathetic. They’re waiting for a software Athena to spring forth fully grown.

The researchers from Exponential Appliances are “not looking for human-level AI; we’re looking for superhuman AI”. And even more to the point: “we aren’t looking for superintelligent employees, we’re looking for
superintelligent products”. Blue Gamma’s digients are clearly not suitable for such a purpose. They aren’t superintelligent, and are unlikely ever to become so, no matter how long their education continues. Their virtual bodies, and their emotions, get in the way of optimizing performance. And even worse (from Exponential Appliances’ point of view), the Blue Gamma digients “think of themselves as persons”, which means that they cannot be treated just as objects, or as commodities. The people from Exponential “want something that responds like a person, but isn’t owed the same obligations as a person”. Ana and Derek, from their years of experience with the digients, know that this is a self-contradictory demand, and that therefore it is impossible. As Ana reflects,

The years she spent raising Jax didn’t just make him fun to talk to, didn’t just provide him with hobbies and a sense of humor. It was what gave him all the attributes Exponential was looking for: fluency at navigating the real world, creativity at solving new problems, judgment you could entrust an important decision to. Every quality that made a person more valuable than a database was a product of experience.

The arguments of the Exponential Appliances researchers remind me of those of David Levy in his book *Love + Sex With Robots* (2007). Levy proposes, on the one hand, that in the near future robots will be advanced enough that they will be entirely indistinguishable from human beings in sexual relationships. They will give their human partners just as much, physically and emotionally, that human lovers do. However, at the same time Levy also presents as an advantage the fact that robots – unlike actual human beings – are infinitely programmable, so they can be guaranteed never to have desires that differ from what their owners want. Therefore,

you don’t have to buy [a robot] endless meals or drinks, take it to the movies or on vacation to romantic but expensive destinations. It will expect nothing from you, no long-term (or even short-term) emotional returns, unless you have chosen it to be programmed to do so.
This is clearly a fantasy (in the most pejorative sense of that term). Levy wants to have things both ways. The robots cannot be both entirely like us, and yet utterly subordinated to our will. If they do our bidding entirely, then they will not seem autonomously intelligent, and they probably won’t even be self-conscious; we will never be able to forget that they are not human. On the other hand, if these sexbots are really as similar to human partners as Levy claims, then they will need to have a degree of autonomy such that we will not be able to completely program them.

It is worth pointing out just how unusual Chiang’s gradualist and experientially-based vision of artificial intelligence is. Science fiction, futurist speculation, and analytic philosophy alike tend either to deny that strong AI is possible at all, or else to present it in apocalyptic terms. John Searle, with his famous “Chinese Room” argument, exemplifies the former alternative. For Searle, mental intentionality “is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena”. Because of this dependency, intelligence cannot be emulated in software. “No [computer] program by itself is sufficient for thinking”, Searle argues, because no software can make the leap from syntactic rules to semantic content. Evidently Searle would reject the very idea that the sort of extrapolation that takes place in Chiang’s story is possible.

At the other extreme, the futurist Ray Kurzweil is fully confident that he will soon be able to download his mind into the computer network, and thereby live forever. He claims that “the AI revolution is the most profound transformation that human civilization will experience”, and that it will inevitably take place before the middle of the 21st century. For Kurzweil, the development of general artificial intelligence will lead to a massive break in human history, a Singularity after which everything in the world will be totally transformed – and supposedly for the better. Meanwhile, scientists like Stephen Hawking, entrepreneurs like Elon Musk, and philosophers like Nick Bostrom have all issued warnings that intelligent machines might well be a threat to humanity. Their “superintelligence” will extend so far beyond ours, Bostrom says, that we will never be able to understand them, let alone control them. Their unchecked growth will menace us with extinction: “once humans develop artificial intelligence it would take off on its own, and re-design itself at an ever increasing rate…”
This is quite possibly the most important and most daunting challenge humanity has ever faced”.

“The Lifecycle of Software Objects” doesn’t entirely rule out the possibility of machine superintelligence. But it implicitly suggests that the Neuroblast approach is much more likely to succeed in creating any sort of machine-based mind. The problem with the vision of superintelligence is that there is really no way to extrapolate it from what exists today. Kurzweil, of course, claims to be doing just this, but his presuppositions and his account of mind are too simplistic to be convincing. Kurzweil explicitly claims, for instance, that the Singularity “is the inexorable result” of Moore’s Law. As computing power becomes steadily cheaper, we will eventually make computers with as many connections as the number of synapses in the human brain. And Kurzweil simply waves away the question of structure and organization. Once this quantitative equality is achieved, he thinks, the rest will just automatically follow.

These visions of superintelligence, whether for good or for ill, derive from an overly grandiose and inflated view of what cognition and consciousness actually are, and what they are actually able to do. We tend to be self-congratulatory about our own cognitive powers. We tell ourselves that our “sapience” is vastly superior to the “mere sentience” of all other organisms. And we also usually privilege our capacity for abstraction, and ignore the ways that our own mentality (like that of other entities, organic or machinic) is emotively based, embodied, and situational. All this provides the ground for our picture of ultra-rational and superintelligent posthuman machines. In short, we imagine “intelligence” to be a kind of irresistible comic-book superpower; and we project a posthuman future on that basis. Thus Kurzweil maintains that “intelligence is more powerful than physics”, able to “maneuver and control” all material forces, and thereby “engineer the universe it wants”. This is a fantasy vision of intellect: without limits, without finitude, without contextual grounding, and without friction.

In Chiang’s novella, in contrast, there is no Singularity, and no real prospect of superintelligence. The digients’ cognitive powers are neither special-purpose, nor vast beyond measure. The novella extrapolates its vision from the state of currently-existing software; and, in just the same way, the digients within the story develop their powers by in effect extrapolating from an already-existing base of software performance. By
grasping mentality in this way, Chiang has no need to posit – as Kurzweil and Bostrom both do – that the development of intelligent software must take place at an ever-expanding, exponential rate, and end by leaving us behind.

This also means that virtual existence will not be as different from physical existence as we sometimes imagine. For both sorts of existence require a certain degree of embodiment. Human intelligence is not just located in our brains; it also necessarily involves some degree of extension into the outer environment, in the form of what David Chalmers and Andy Clark call the “extended mind”. It is therefore impossible to disentangle biological intelligence from its “artificial” prosthetics and extensions – which range, in the case of human beings, all the way from drawing pictures in the sand, to writing technologies, to the latest computational innovations.

This is yet another reason why Kurzweil’s vision of infinite intelligence is ludicrous. We may well develop sophisticated, radically nonhuman forms of intelligence; but the necessity for embodiment and extension will mean that this intelligence is still subject to biological and energetic limitations, or – as with Chiang’s digients – to virtual equivalents of these. Kurzweil imagines that downloading his mind into the network will free it from all physical constraints. But even if he succeeds in this task, he is bound to be disappointed. Since physical and sensory interactions, embodied feedback mechanisms, and other extensions into the environment are crucial parts of mental functioning, Kurzweil will need to take them all along with him when he disperses into the network.

In fact, “The Lifecycle of Software Objects” presents us with an exact inversion of Kurzweil’s scenario. Instead of human beings downloading their minds into the network, we get the digients temporarily “uploading” themselves into the physical world. This is made possible by the manufacture of

a robot body, newly arrived from the fabrication facility. The robot is humanoid in shape but small, less than three feet in height, to keep the inertia of its limbs low and allow it a moderate amount of agility. Its skin is glossy black and its head is disproportionately large, with a surface mostly occupied by a wraparound display screen.
The digient software is simply redirected, so that it controls this physical body instead of its usual virtual one. Jax is able to do this easily, “because the test avatar isn’t radically different from his own; it’s bulkier, but the limbs and torso have similar proportions”. Instead of seeing, hearing, and feeling its virtual body and virtual surroundings, the digient is able to feel a sense of presence in the actual physical world, thanks to the robot unit’s cameras, microphones, and “tactile sensors”. All this is something of a trick, but it works well enough: Ana “knows that [Jax is] not really in the body – Jax’s code is still being run on the network, and this robot is just a fancy peripheral – but the illusion is perfect”.

Indeed, it isn’t just that the Neuroblast digients have a limited and bounded intelligence. It is also that they are too autonomous, and too playful, to work for businesses or individuals as digital assistants. They don’t like tedious and repetitious jobs any more than human beings do. If the point is to automate such jobs, so that conscious beings don’t have to endure them, then developing self-conscious AIs is economically and ethically counterproductive. Generalizing from the results of Neuroblast and other competing programs, “many technology pundits declare digients to be a dead end, proof that embodied AI is useless for anything beyond entertainment”.

All this changes with “the introduction of a new genomic engine called Sophonce. Where the Neuroblast digients are programmed to be “adorable”, so that owners will bond with them, the ones produced by Sophonce are single-minded, task-oriented, unsympathetic, and utterly charmless. The Sophonce digients are deliberately manufactured so as to exhibit “asocial behavior and obsessive personalities”, ideally suited to business contexts. Since they do not need to develop conviviality, playfulness, and other social skills, they do not require anywhere as much human intervention and training as the Neuroblast digients do. The problem with them is that they are so unengaging “that few people want to engage in even the limited amounts of interaction that the digients require”.

Enter a company called Polytope. It is trying to produce a new breed of smart digital assistants. The plan is to augment the capabilities of the Sophonce digients by hands-on training of the sort that Ana provided to Jax and other Blue Gamma/Neuroblast digients. This would supposedly give the Polytope’s digients the best of both worlds. The problem, of course, is that the Sophonce digients cannot establish emotional relationships with
their trainers, or with anyone. The Polytope people hope to get around this by requiring the human trainers to use something called InstantRapport: “one of the smart transdermals, a patch that delivers doses of an oxytocin-opioid cocktail whenever the wearer is in the presence of a specific person”. The company reasons that “the only way trainers will feel affection for Sophonce digients is with pharmaceutical intervention”. The digients will not ever feel any sort of empathy on their end, but the human trainers will forcibly develop affection for, and empathize with, those digients nonetheless.

Ana is tempted to take the job, despite its creepiness, if in return Polytope agrees to port the Neuroblast digients to Real Space. But of course, this raises a whole set of questions. Will Ana still be able to care for Jax in the same way, once she starts spending most of her time with a Sophonce/Polytope digient instead? More generally, what does it mean to freely agree to a procedure that changes the very basis of who you are, in a way that you cannot control? Moreover, what does it mean to do this as a condition of employment? The situation is quite different from voluntary procedures like getting plastic surgery, or LASIK eye surgery, or (in the near future) intelligence augmentation through chemical or genetic means. It may well be that chemical intervention can change our personalities, in the same way that rewriting code can change the personalities of the digients. But the question of consent, both for human beings and for digients, remains murky and troublesome.

The other alternative for rescuing Ana’s and Derek’s digients is equally creepy. A company called Binary Desire will gladly port the Neuroblast code to Real Space, in return for being allowed to license the digients as sexbots. The Binary Desire people emphasize that they are actively trying to get away from cheap and sleazy exploitation. “As long as there have been digients, there have been people trying to have sex with them”; but this has usually happened at a very low level. In the world of the novella, there are already digital entities like Sophonce digients “dressed in Marilyn Monroe avatars, all bleating Wanna suck dick. It’s not pretty.”

Binary Desire seeks instead to develop virtual “sex partners with real personality”:

As the digient gets to know a human, we’ll enhance the emotional dimension of their interactions, both sexual and non-sexual, so
they’ll generate love in the digient… For the digient, it will be indistinguishable from falling in love spontaneously.

Also, Binary Desire promises to “retain the circuit-breakers” that prevent the digients from feeling pain, so that they will never become the victims of sadists:

The digients won’t be subjected to any coercion, not even economic coercion. If we wanted to sell faked sexual desire, there are cheaper ways we could do it. The whole point of this enterprise is to create an alternative to fake desire. We believe that sex is better when both parties enjoy it; better as an experience, and better for society.

The playful and emotional nature of the Blue Gamma digients makes them perfect candidates for the Binary Desire plan. Ana notes that it is a bit “like a Neuroblast version of Instant-Rapport”. The difference is that here it is the digients’ personalities that are manipulated, rather than those of the human trainers and partners. If digients are programmed this way, they won’t have “any choice about what they enjoy”. But the Binary Desire people deny that the situation is “any different for humans… We become sexual beings whether we want to or not”. It’s just that biological human beings are programmed, or re-programmed, by electrochemical changes, instead of by rewritten code.

“The Lifecycle of Software Objects” doesn’t offer an answer to any of these dilemmas. There is no way to resolve the ageold debate between free will and determinism, for instance. The novella instead suggests that whatever is true for us must also be true for the digients. To the extent that we can make spontaneous, unforced decisions, so can they. And to the extent that the digients are susceptible to being manipulated from outside, the same is true for biological human beings. This is not changed, in principle, by the fact that we have access to the digients’ source code, but not to our own – regardless of what, if anything, the biological equivalent of “source code” might turn out to be.

Chiang’s novella also makes the point that technological developments can never be separated from social and economic ones. No research program can be pursued without sufficient funding. In our current neoliberal climate, this means that the development of artificial intelligence
is necessarily subject to corporate control, and can only be pursued if, and to the extent that, it promises profit. (The one exception to this, not discussed in the story, is research conducted secretly by the military and the security services.) Ultimately, play and pleasure – the initial endowment of the Neuroblast/Blue Gamma digients, and the reason why people like Ana and Derek are so attached to them – must be subordinated to economic considerations. Blue Gamma goes out of business, and the three alternatives Ana and Derek must face in order to keep their digients going all involve monetizing, and restrictively channeling, the digients’ abilities.

Along these lines, it is ironic, but not particularly surprising, that the only way to give the digients legal rights – to endow them with any degree of autonomy, or with the legal status of personhood – is to register them as corporations:

Artificial-life hobbyists all agree on the impossibility of digients ever getting legal protection as a class, citing dogs as an example: human compassion for dogs is both deep and wide, but the euthanasia of dogs in pet shelters amounts to an ongoing canine holocaust, and if the courts haven’t put a stop to that, they certainly aren’t going to grant protection to entities that lack a heartbeat. Given this, some owners believe the most they can hope for is legal protection on an individual basis: by filing articles of incorporation on a specific digient, an owner can take advantage of a substantial body of case law that establishes rights for nonhuman entities.

This makes a grim sort of sense when we consider that corporations are not only recognized as “persons” by the United States courts, but even granted freedoms and rights that biological persons do not enjoy. At one point in the novella, Marco and Polo ask Derek if he will register them as corporations, so that they “can do whatever [they] want”. At another point, Jax asks Ana if he can get a job, so that he will be able to pay for her to continue taking care of him. The digients’ dependence on their human trainers, and their prospective independence from those trainers, are both financially mediated in the long term.

The largest tension running through the novella is that between the digients’ sheer existence and their economic utility. Where Sophonce digients have particular marketable skills, the Blue Gamma/Neuroblast
digients are characterized, above all, by their playfulness and curiosity. They exhibit what Alfred North Whitehead calls “a certain absoluteness of self-enjoyment”. Their sentience is far more a matter of feeling, than it is one of cognition. And this is why their existence is so precarious. The things that they do are gratuitous rather than functional, which means that – short of turning them into sexbots – they cannot really be monetized. Most recent philosophical accounts of mind are entirely functionalist and cognitivist. Feeling and emotion only play secondary roles. As Robert Zajonc summarizes it, for cognitivism “affect cannot be independent of cognition because by definition cognition is a necessary precondition for affective arousal”. But “The Lifecycle of Software Objects” insists that this is wrong. The cognitive skills of the Neuroblast digients are secondary to their emotions. If “experience is the best teacher”, this is because it is only through the adventures of affect that Jax and the other digients are able to learn to perform cognitive tasks in the first place. If they can walk, talk, read, and otherwise evaluate and negotiate their way through their environment, it’s because they already have a certain basic sensitivity. The analytic-philosophical privileging of cognition over affect is of a piece with the economic privileging of the digients’ business skills (or for that matter, sexual skills) over their own self-enjoyment. “The Lifecycle of Software Objects” doesn’t suggest that we can ever escape these sorts of constraints, but it does tell us that they aren’t the last word.
FOUR

Thinking Like a Human Being

Scott Bakker’s novel *Neuropath* (2008) is best described as a near-future science fiction thriller. It takes place in a world that is recognizably our own, only a few years down the line. Everyday life for affluent North Americans is not much different in the future world of the novel from how it is today; but enough time has passed that political and environmental conditions have gotten significantly worse. On the TV news, we hear about such events as terrorist attacks on Moscow and the American Southwest, “French ecoriots”, and “the Chinese economic crisis”. The news also features “a tasteless story about Ray Kurzweil’s recent death”: evidently the futurist does not achieve his goal of attaining immortality by uploading his consciousness onto the Internet. More significantly, we learn of “the emergency repeal of the constitutional provisions guaranteeing due process” under United States law. In fear of terrorism, we are told, “the American public had enthusiastically surrendered their constitutional scruples”. In this near-future America, the National Security Agency (NSA) can permanently detain anyone they choose. Either they arrest you directly; or else they plant child pornography on your computer, and let the local authorities do the job. “Scarcely a month passed”, the novel tells us, “without some story of some reform-minded political figure arrested on child pornography charges”.

In this setting, *Neuropath* tells the story of Thomas Bible, a forty-something psychology professor, divorced, with two children and a dog, living a humdrum life in the “packed anonymity” of the suburbs. Thomas is the author of a book, *Through the Brain Darkly*, which presents a revolutionary new theory of the human mind. The book is too extreme to have made much of a public impact: “the reviews were harsh; it went out of print”. Yet everything in the novel turns upon Thomas’ theory, which is referred to, throughout the book, as “the Argument” (with a capital A). *Neuropath* is science fiction, not just because it is set slightly in the future, but more importantly because the Argument that it presents is extrapolated, as an Author’s Note points out, from “actual trends and discoveries in
neuroscience, psychology, and cognitive science”. These trends and discoveries all point to the disturbing truth that “we are not what we think we are”.

The Argument starts from the observation that science always understands things “in terms of quantity and function instead of quality and intention”. Indeed, “wherever science encounters intention or purpose in the world, it snuffs it out. The world as described by science is arbitrary and random. There’s innumerable causes for everything, but no reasons for anything”. Physical science is a war machine, a weapon of mass destruction. As a result of its relentless progress over the past several centuries, “science has pretty much scrubbed psychology from the natural world”. We no longer believe that natural events carry omens and convey messages. We use weather satellites to track oncoming storms, instead of blaming Thor for the thunder.

This aspect of Neuropath ’s Argument accords well with the naturalistic currents in contemporary philosophy. The speculative realist philosopher Ray Brassier, drawing upon both continental and analytic sources, makes a similar point. Brassier insists that, due to the success of physical science over the past few centuries, the “intelligibility” of the world has become detached from meaning: with modern science, conceptual rationality weans itself from the narrative structures that continue to prevail in theology and theologically inflected metaphysics… The world has no author and there is no story enciphered in the structure of reality. No narrative is unfolding in nature.

Such an observation is scarcely even controversial any longer. Weber’s thesis on the disenchantment of the world has become our obvious condition. “Contemporary culture”, as we are told in Neuropath, has long since “digested themeaninglessness of natural events, the fact that they [are] indifferent to all things human”.

But what happens when we apply the scientific method, not just to the surrounding physical world, but also to ourselves, and especially to our minds? This is something that we still find disquieting. As the Argument insists, if naturalism is right, then what holds true for all other entities in the universe must hold true for human beings as well. The things that we have
come to understand about the world must be applied as well to our own processes of understanding. And indeed, recent advances in neuroscience and cognitive psychology have taught us a lot about the brain. This is largely due to new technologies, like fMRI (functional magnetic resonance imaging), first developed in 1992, which allows us to track brain activity in real time; and TMS (transcranial magnetic stimulation), first successfully used in 1985, which allows us to affect targeted portions of the brain in such a way as to alter a person’s feelings, attitudes, and judgments. Through these techniques, together with advances in computing power, we have arguably learned more about the physical functioning of the brain in the past thirty years or so, than we did in all of previous human history.

Of course, the relation of brain to mind, or of electro-chemical processes in our neurons to full-fledged subjective experience, is still hugely controversial. But Neuropath’s Argument takes it as a given that science is now in process of scrubbing psychology from the human world, just as it previously scrubbed psychology from the natural world. That is to say, the psyche itself is rapidly being depsychologized, as paradoxical as this might sound. Even the decenterings of subjectivity proposed by psychoanalysis and deconstruction have not really prepared us for this eventuality. We cannot help believing that we have reasons for what we do; “human beings explain and understand themselves in terms of intentions, desires, purposes, hopes, and so on”. But in fact, as Neuropath reminds us, “every thought, every experience, every element of your consciousness is a product of various neural processes”. You might think that you have actively chosen to do something or other; but “as a matter of fact – fact, unfortunately, not speculation – your brain simply processed a chain of sensory inputs… then generated a particular behavioral output”. The brain is “a machine that generates behaviors which are either repeated or not depending on how the resulting environmental feedback stimulates its pleasure or pain systems”. We like to imagine ourselves as free, rational beings; but actually (to quote the novel at its most harshly cynical) “we’re simply meat puppets deluded into believing we live in a moral and meaningful world”.

Brassier makes a similar point: today, he says, it is no longer possible, as Nietzsche suggested in the 19th century, and as the existentialists still maintained in the mid 20th, “for human consciousness to provide the meaning that was absent from nature”. This is because we can no longer grant any special status to our own subjectivity. “The meanings generated
by consciousness can themselves be understood and explained as the products of purposeless but perfectly intelligible processes, which are at once neurobiological and sociohistorical”. Brassier warns us that human subjectivity no longer provides a refuge from science’s relentless demystification and disenchantment of the world.

In fact, *Neuropath* ’s Argument tells us that things are even more extreme than this. Recent research in neurobiology and cognitive psychology shows that most of the neural processes that go on in our brains are not consciously accessible to us at all. This is why our actual self-awareness is so misleading, incomplete, and prone to illusion and error. This is more thoroughly the case than Freud ever imagined. Our attempts at self-examination through introspection are incompetent at best, and delusional at worst. Most of what we think about ourselves is biased and inaccurate. Pushing the negative results of this recent research as far and as provocatively as possible, the Argument in *Neuropath* suggests that the human mind is altogether incapable – in principle, and not just in fact – of understanding itself. Things like intentions, meanings, and purposes “only seem real because we’re riding the neural horse backward”. But since we remain necessarily committed to such artifacts, we remain inescapably blind to the actual material causes of our thoughts and actions. We are constitutively unable to trace our own mental states back to the electrochemical events in the brain that produce them.

Our psychological self-explanations are therefore best understood as fictional narratives, or confabulations. We may think here of Benjamin Libet’s famous experiments on “the precursors of decision-making”. These experiments are frequently referenced both in affect theory and the philosophy of mind; and they are explicitly referred to at one point in the novel. Libet discovered that “readiness potentials” for a given action build up in the neurons of a subject’s brain, half of a second before the subject consciously decides to perform that action. In effect, our decisions have already been made for us – or at the very least made by nonconscious processes within us – prior to our very awareness of making them. Here the third person trumps the first person: mental events that are inaccessible to introspection can nonetheless be measured and recorded by objectifying scientific instruments. Under such conditions, my sense of exercising “free will” is only a self-deluding attempt to give myself credit for an event in my brain that has already happened. In the words of the novel, “willing… is an
add-on of some kind, something that comes to us after the fact”. It is hard for us to fully grasp “just how after the fact conscious experience is”. This temporal lag is crucial. It means that our minds are never able to keep up with themselves. The media theorist Mark Hansen has insisted that human experience is currently undergoing a fundamental transformation caused by the complex entanglement of humans within networks of media technologies that operate predominantly, if not almost entirely, outside the scope of human modes of awareness (consciousness, attention, sense perception, etc).

The Argument in Neuropath suggests that this is not only the case in relation to computerized microsensors and other forms of what Hansen calls “21st century media”, but applies globally in terms of the brain’s relation to itself. “The brain simply isn’t equipped to keep track of itself… it lacks the processing power… The best it can do is scribble cartoons of itself”. The Argument tells us that everyday life, as we experience it, is nothing more than such a “cartoon”. We live, inescapably, in “Disney World”: a world that is “papered over with conceit after comforting conceit… anchored in psychological need rather than physical fact”. This is inevitable, because the bulk of your brain’s processing falls outside what you can experience; it simply doesn’t exist for your consciousness, not even as an absence… Our brains… are entirely blind to the deep processing that drives them… The neural correlates of consciousness have no access to the real neurophysiological movers and shakers down below.

In this regard, we are even worse off than the mythological prisoners in the cave of Plato’s Republic. For they at least have the faint hope of becoming aware of their plight, and ascending towards the light. But we are never released from our illusions, not even when we become aware that our experiences are illusions. We are never able to see “the shadow behind the occluded frame”. That is to say, since the “frame” that limits my awareness is itself “occluded”, I am unable to realize that my experience is, in point of fact, circumscribed and partial. Since I cannot perceive the boundaries of my experience, I cannot even grasp that my experience is limited, rather
than being comprehensive. This is why, for instance as the novel observes
my “visual field” seems to “simply ‘run out’ without having any visible
edge”.

The Argument also undermines our common intuitions about our own
inner sensations, or what the philosophers call *qualia*. I tend to think, as
many people do, that there is a certain vividness and intensity to my inner
life. But this *qualitative* dimension of my experience is something that I
cannot capture and put into words. For instance, I can tell you that I am
seeing a particular shade of the color red; but I do not know how to convey
to you – or even describe to myself – the deep, and quite specific, *feeling* of
seeing this precise shade of red. I don’t even know how to explain its
difference from other closely related, but ever so slightly distinct, shades of
red. I am inclined to presume that, so long as you are not blind or
colorblind, your experience of this particular shade is similar to mine. But I
cannot ever prove this, as there is no way of actually comparing my inner
experience with yours. This is the difficulty that Wittgenstein struggles with
in his *Philosophical Investigations*, and that underlies Thomas Nagel’s
meditation on the difficulty of knowing what it is like to be a bat.

The problem would seem to be that there is a mismatch between the
functional information processing that goes on in the brain, on the one
hand, and the phenomenal experience that emerges as a result of this
processing, on the other. The former is accessible to objective, third-person
measurement with things like fMRI scans; but the latter is not. Many
philosophers of mind – David Chalmers is one prominent example – seize
upon this imbalance, and maintain that information processing is simply *not
enough* to account for the “rich inner life” of conscious entities. This is
why, according to Chalmers, subjective experience itself is not “susceptible
to the standard methods of cognitive science, whereby a phenomenon is
explained in terms of computational or neural mechanisms”. On the other
side, reductionist thinkers – most famously Daniel Dennett – respond that
this seeming mismatch is itself only an illusion. If we look carefully enough
at the phenomena of consciousness, Dennett says, we discover that all of its
features do in fact reduce, without remainder, to computational and neural
mechanisms. Dennett therefore maintains that “consciousness cannot be
separated from function”, and that there are no “absolutely indescribable
properties in our experience”.

Now, Scott Bakker is evidently much closer to Dennett’s reductionist view of consciousness than he is to Chalmers’ expansionary one. But the Argument in *Neuropath* takes a radical step that Dennett does not, and thereby turns the whole dispute about *qualia* inside out. Instead of denying, as Dennett does, the very claim of a mismatch between information processing and inner, phenomenal experience, the Argument in *Neuropath* accepts the existence of a disparity, only to reverse its terms. For the Argument suggests that phenomenal experience is the consequence, not of an overflowing qualitative richness, but rather of a fundamental *deficiency*. We see, hear, and feel in the way that we do as a result of informational impoverishment. No matter what I experience, “the neural processing that makes these experiences possible… is *utterly invisible*” in and of itself. What seem to be the *positive* features of conscious experience are therefore really symptoms of this basic deficiency, the negative consequences of limitation and ignorance. My experience of this particular shade of red is ineffable, not because it exceeds any sort of measure, but because it is too “subtle” (as Thomas Metzinger puts it) to reach the threshold of discernibility needed for even the most minimal cognition.

According to the Argument, what holds for the qualitative nature of consciousness holds also for its other seemingly inexplicable features, such as its unity and presentness. “Experiences are always unitary, and they’re always now”, the Argument tells us, because “they’re byproducts of what the brain lacks”. Our minds perpetually suffer from “things like inattentional blindness, change blindness, masking, perceptual asynchrony, processing lags, and so on… You could make a career out of cataloging all the ways in which consciousness is either blinkered or outright deceptive… Out of all the information our brains crunch every second, only a tiny sliver makes it to conscious experience – less than a millionth, by some estimates”. The seeming richness of my first-person perspective is a hallucinatory effect of this fundamental sparseness.

What does this mean for human subjective experience? We live in “a world of pure experience”, William James told us long ago. In the words of the novel, we cannot help believing that experience is “pure and bone-deep. What could be more true than that? What could be more true than the feelings that underwrite our very existence?” And yet, the latest scientific evidence suggests that we do not even “experience experience *as it is*”. Rather, experience itself is “profoundly deceptive”. The Argument tells us
that this discordance is precisely as “we should expect”, once we have learned how untrustworthy our own beliefs can be, and how mistaken our introspection.

I think that it is worth insisting upon the sheer, radical outrageousness of this claim. Modern philosophy, from Descartes all the way through to phenomenology, is grounded upon the self-evident givenness of immediate experience, and subsequently of our introspective reflection upon this experience. Descartes imagines an “evil genius, supremely powerful and clever, who has directed his entire effort at deceiving me”. Yet even if all the things I think are false, Descartes says, it still remains true that I exist, because I am the one who is thinking these things.

Recent cognitive psychology, however, casts doubt even upon this minimal self-assurance. For apparently experience still happens even when there is no hope of attributing it to myself. The Argument therefore draws a wedge between experience itself, and any hope that I might have of ‘owning it’, let alone reflecting upon it clearly. I can easily be mistaken when I grasp my own experience; and in any case, I can only lay hold of it after the fact. “You like to think that you have all these experiences, that you author all of your actions, but the sad fact, my dear, is that you simply accompany them”. Instead of Descartes’ “I think, therefore I am”, the Argument suggests that “something like ‘it thinks, therefore I was ’ would probably be more accurate”.

The Argument, as it is put forth in Neuropath, is a sketch for what Scott Bakker subsequently develops in a fuller and more explicit form on his blog, Three Pound Brain. Here, he calls it the Blind Brain Theory (BBT). Indeed, at one point in Neuropath, the Argument is formally referred to as the “Blind Brain Hypothesis”. In moving from a hypothesis to a full-blown theory, the blog marshals experimental evidence, and engages in straightforward philosophical disputation, in a way that the novel does not. The blog seeks to rigorously establish what the novel merely asserts. In Three Pound Brain, Bakker argues for the truth of the BBT; he elaborates and complicates the theory, shows how it addresses and resolves various conceptual impasses, and contrasts it with other theses in the philosophy of mind. None of this is done in the novel.

However, all this does not mean that Neuropath is just a mere illustration of the BBT. My claim about the novel, as about science fiction more generally, is precisely the opposite. The difference between blog and
novel, as more broadly between philosophy and science fiction, has to do with the way that the latter works as a thought experiment, and an exploration of extreme possibilities. In Neuropath, Bakker does not attempt to prove the Argument, philosophically or empirically. Rather, he explores the (largely horrific) consequences of the Argument: those that would result from its being true, as well as those that might result from people becoming aware of it, maintaining it, and putting it into practice. What is really at stake in Neuropath is the question of how the world is changed – personally, socially, and technologically – once the Argument has become conceivable; and even more, once it has become operational. Human consciousness may well have always been delusional, but recent technological inventions make it possible to mobilize our delusions in a new way.

The ultimate consequence envisioned by Neuropath is what it calls “the semantic apocalypse, the apocalypse of meaning”. This is a conflagration in which our common-sense intuitions about ourselves are discredited once and for all. It’s not just a matter of so-called folk psychology being replaced by a more scientific vocabulary, as some philosophers have imagined. Rather, our very ability to make sense of our own experience is paralyzed. We are no longer able to believe that anything in our lives is meaningful – or even that anything in our lives can be referred back to ourselves. This is not existential alienation, nor even totalitarian mind control, but something worse: an even more extreme self-divestment. In one of the novel’s most lurid passages, Thomas is told: “You want to believe I’m doing things to you, when in fact I’m doing things with you. The only reason I can play your thoughts and experiences like a sock puppet is because that’s what you are”. It’s a rather extreme, and negative, version of what used to be celebrated as Romantic inspiration: “not I, but the wind that blows through me”.

We may compare this, once again, to Brassier’s claim that any effort “to wrest some sort of psychologically satisfying narrative from elements of the modern scientific worldview… is doomed because it is the very category of narrative that has been rendered cognitively redundant by modern science”. In a certain sense, Neuropath tells us, the semantic apocalypse has "already happened". It is too late for us to retreat from the implications of the new neuroscience and cognitive science. For the first time in human history, perhaps, we are no longer able to escape or deny – as previous generations
did – “the nihilistic truth of existence”. We can no longer tell ourselves the kinds of stories that make our lives meaningful again.

I think that there is another way to say this, though it is one that the novel does not put forth explicitly. Consider the history of modern, scientific conceptualizations of the mind. These generally track the most advanced technologies available at any given time. Materialists of the 18th century speculated that the mind worked like a clockwork mechanism. In the 19th and earlier 20th centuries, Freud was only one of many theorists who modeled the mind instead in thermodynamic terms, as if it were something like a steam engine or a vast hydraulic system. In the later 20th century, thanks to cybernetics and the development of computing, the mind was conceived in terms of information processing. It was thought to be like a digital computer, with the brain as hardware, and mental processes as platform-independent software.

Today, in the early 21st century, we still take for granted (excessively, in my view) the primacy of information processing. But we are starting to abandon the idea of mental happenings as immaterial, platform-independent patterns. Ray Kurzweil’s fantasy of uploading his consciousness to a computer can no longer be taken seriously. Instead, we now seek to explain the mind directly in terms of the electro-chemical processes that actually occur within the brain. With fMRI and TMS, we have moved from technologies that serve as metaphors for the mind, to technologies that themselves literally act upon the mind, by measuring the flow of blood in the brain, and by stimulating or inhibiting particular neurons in determinate ways.

In other words, the question of the mind today has become more a matter of engineering practice, than one of scientific and philosophical understanding. The Argument is less a matter of what the mind is, and of what we can know about it, than of what the mind does, and what we can in turn do with it or to it. The shift from psychological reasons and intentions to physiological causes and mechanisms is also a shift from epistemology to instrumentality.

This is why the Argument needs to be addressed pragmatically, in terms of its consequences, more than it needs to be confirmed or discredited through philosophical argument. And such an exploration of consequences, as I have already suggested, is more the province of science fictional speculation and extrapolation, than it is one of strictly philosophical
discussion. Brassier claims that “the very category of narrative” has become “cognitively redundant” as a result of recent neurobiological findings, but *Neuropath* demonstrates that even this redundancy still needs to be addressed in narrative.

The Argument’s effects and consequences are dramatized in *Neuropath* in terms of what can best be described as an ongoing duel between the two main characters: Thomas Bible, the original author of the Argument, and his erstwhile best friend Neil Cassidy, a neurosurgeon who puts the Argument into actual practice. Thomas and Neil have a long relationship, dating back to when they were college roommates. Indeed, Thomas originally develops the Argument as an undergraduate, in the course of “college bull sessions” with Neil. The novel describes these sessions as an experience of mutual infatuation and intoxication: like taking “a kind of experimental drug”, or even like “a religious experience”. Even science-based demystification tends to be grasped in mystical or intoxicating terms.

As Thomas and Neil elaborate the Argument, it both frightens and excites them. With other people, they use it as a sort of social weapon, trotting it out for things like “mopping the floor with lit majors, freaking people out around the bong”. For simply to maintain the Argument is to display your bravado. It shows that you are tough enough to handle its nihilistic implications – unlike all your shocked and wimpy listeners, who cannot accept it because it “cut[s] against the grain of too much hardwiring and socialization”. The Argument is necessarily a provocation, and thereby a powerful tool in a game of macho one-upmanship. And this continues to be the case throughout *Neuropath*. Even at the end of the book, when he is being tortured by Neil, who claims to have pushed the implications of the Argument much further than Thomas has ever dared – even then, Thomas still admires and looks up to the alpha-male exuberance of his friend-turned-tormentor. Thomas cannot help wondering, “where did [Neil] find the *balls* to do the things he did?” The very outrageousness of the Argument becomes a reason for espousing it, and a mark of the power of whoever does.

In the present time of the novel, Neil shows up on Thomas’ doorstep in order to, in effect, reproach him for backsliding from the rigor of maintaining the Argument. “Like most, Thomas had moved on… The years passed, the children grew, and he found himself packing all the old questions away, even as he continued playing Professor Bible, destroyer of
worlds in the classroom. Nothing killed old revelations quite so effectively as responsibility and routine”. Thomas still maintains the Argument in theory, but he doesn’t imagine that people in general are "capable of believing it. And he doesn’t himself live as if it were true. His humdrum existence as a suburban, divorced father-of-two continues unchanged. Thomas is sort of like those 19th century atheists who (much to the scorn of Nietzsche) continued to act in everyday life in the same ways, and with the same morality, as they would have done if they still believed in God.

Neil, in contrast, “had never let go” of the Argument; his entire life consists of following it to its furthest consequences. And this is what sets the narrative of Neuropath into motion. When Neil visits Thomas, he reveals that ever since medical school he has been working for the NSA, putting his expertise to use in “neuromanipulation”. Inspired by the Argument, Neil develops new technologies for interrogating prisoners. This is the science fictional core of Neuropath. Just as the Argument itself is extrapolated from recent scientific discoveries, so Neil’s inventions – inspired by the Argument – are extrapolated from, and projected beyond, what scientists are actually capable of doing today. At least, I hope that the technologies imagined in the book are not yet possible, and have not already been put into use in Langley, Virginia and Guantanamo Bay.

In the near-future world of Neuropath, the NSA has discovered how to manipulate the brains of “terrorists” and other political prisoners. The real problem with torture, from the NSA’s point of view, is that it is not nearly effective enough. Why waste so much time and energy trying to break a prisoner’s will to resist, when we now know that this “will” is “simply one more neural mechanism” that can easily be put “offline”? As Neil explains to Thomas, when interrogating prisoners “we simply isolated the offending circuits and shut them off. It was as easy as flicking a switch”. Once this is done, the prisoner ceases all resistance, and is happy to reveal everything he knows. Mental privacy does not need to be violated, when it can much more easily just be shunted aside and rendered irrelevant. “Why design a machine to read thoughts”, Neil says, “when all you have to do is shut down a few circuits and have your subject read them out for you?”

Horrible as this is, it is only the beginning of the way in which Neil works through the pragmatic consequences of the Argument. The main plot line of Neuropath involves Neil going rogue from the NSA, and engaging on a crime spree that seems designed to demonstrate the truth of the
Argument. Neil kidnaps people and performs gruesome operations upon them in order to alter their minds. The most prominent victim is Theodoros Gyges, a billionaire tycoon. Neil kidnaps Gyges, surgically alters his brain in order to create two specific cognitive impairments, and then releases him again. The first impairment is prosopagnosia, more commonly known as face blindness. This is a rare syndrome, but one that has been much studied, and is well known to scientists and clinicians. Gyges’ sense of vision is not in itself impaired; but he becomes unable to recognize and remember human faces. As a result, his sense of identity shatters. His family and friends come to seem like “faceless monstrosities”, alien intruders. “When I stare at you”, Gyges says, “I don’t recognize your face from one second to the next. And it’s not like your face becomes something new every moment, something that I’ve never seen before. It’s just unknown. Unknowable”.

Ms disorientation also applies to Gyges’ sense of himself; he cannot even recognize his own face in the mirror. Is personal identity still possible without the power of recognition? Gyges’ infirmity works so as “to undermine the notion of personhood” altogether.

Neil also implants a second infirmity in Gyges’ brain: a sexual compulsion that transforms him into a gruesome serial killer, known in the tabloids as the Chiropractor. In this guise, Gyges not only rapes and murders women, but eviscerates their bodies as well. He easily evades detection and capture, despite the efforts of a huge task force to track him. Presumably he cannot be accurately profiled, because he is no longer a “self” in any conventional sense. In the original (UK) version of Neuropath – though not in the slightly revised US version – short first-person passages from the Chiropractor’s perspective, printed in italics, are interspersed between the third-person chapters that tell Thomas’ story. These passages are oddly impersonal, despite being voiced by an I. “Oh yes, I see you”, The Chiropractor imagines saying to one of his victims. “As still as a magazine cover. As blank as a porn star between takes… At long last, you mean only what I want you to mean”. The Chiropractor’s murders are a kind of lurid, pulp version of Neil’s and the NSA’s experiments, wiping the slate clean of the victims’ own intentions and meanings.

In the course of the novel, Neil kidnaps and performs surgery upon a number of other test subjects besides Gyges. These episodes all work as demonstrations of the Argument. Neil rewires the brain of a porn actress so that she literally orgasms while mutilating herself to death. He reprograms a
Congressman who frequently pontificates about free will and responsibility, so that the man takes bites out of the flesh of a living 10-year-old girl, all the while proclaiming that he does not want to do this. Neil then subjects a televangelist to alternate spasms of feeling himself to be secure in God’s grace and knowing himself to have been eternally damned to Hell. Through all of Neil’s actions, the Argument – as Thomas comes to recognize – is “not simply paraphrased, but enacted”.

Such assaults, Neil tells Thomas, are “meant to get your brain processing the Argument again, to reacquaint you in the most urgent and intimate way with the force of your own logic”. And they are followed by the climactic confrontation of the novel, in which Neil straps Thomas into the Marionette device, known colloquially as “Mary”: an NSA contraption which is able to induce whatever mood or feeling the operator wishes to implant in the subject’s mind. This machine is, once again, only a slight extrapolation from actually existing technology. It is “adapted”, Neil tells Thomas, “from stereotactic neuroradial surgical devices – you know, the ones that use overlapping particle beams to burn out tumors? We found a way of doping the blood so that we could exercise pinpoint metabolic control at multiple points in the brain”.

With Mary, Neil leads Thomas through a roller coaster ride of different mental experiences, all of which feel as “real” as any experience ever does. Thomas is made to feel “ambient well-being” one moment, then panic and dread the next. He cycles through such states as orgasmic release, out-of-body experience, the collapse of the visual field and warping of “extrapersonal space”, the impression that what Neil is telling him is being spoken by his own voice, and the sense that, “no matter what you’re looking at, you’re convinced that you’re willing it to happen”. Thomas is filled with pain, and then given the sense that “a strange buoyancy filled everything, made candy of all the sharp edges. It suddenly seemed that he watched a rubber world, a place filled with foam simulacra”.

But what can it mean for Thomas to discover, through these experiences, that “Neil had transformed him into the demonstration of his own outrageous claim”? The Argument, in all its extremity, may well be true. But can it ever be entirely believed? I may, at best, accept it intellectually; but how could I espouse it with full conviction? For if I were to do this, I would be forced to recognize that my own “convictions” and beliefs – including my belief in the Argument – are themselves nothing but fictions.
Such is the central enigma around which Neuropath turns. In order to truly grasp the Argument, you cannot just assert it; you have to actually live it and experience it. But this can only happen if you are forced to embrace it, as Thomas is forced by means of the Marionette machine. For the consequences of the Argument are intrinsically inaccessible to experience – and indeed, strictly speaking, unlivable. In short, the Argument cannot be adopted existentially. It can only be enacted, instrumentalized, or put into effect.

In philosophical terms, The Argument involves, and indeed requires, a performative contradiction. It’s a bit like saying, “this statement is a lie”, or “everything is relative”. For if The Argument is true, then logically speaking I cannot possibly have the credibility, or the authority, to affirm its truth. After all, as Thomas puts it at one point, “everyone thinks they’ve won the Magical Belief Lottery… Everyone thinks they more or less have a handle on things, that they, as opposed to the billions who disagree with them, have somehow lucked into the one true belief system”. How could anyone who believes The Argument be any different, when The Argument itself makes clear that he or she is not?

This observation does not in itself refute the Argument. It’s little more than a cheap debating trick to say, for instance, that relativism cannot possibly be true, because if you say that everything is relative, then you are making a statement that itself isn’t relative, but absolute. Such a criticism ignores the way that statements (including argumentative and rationalistic ones) are themselves always situational and mediated. Relativists, like advocates of the Argument, may well discredit themselves by the manner in which they make their claims, but this does not necessarily undermine the truth of their assertions. The Argument suggests that everyone believes that they have won the Magical Belief Lottery, but that nobody has any good grounds for their belief. The fact that this is also the psychological case for somebody who believes the Argument is not a refutation of the Argument, but yet another piece of evidence in its favor.

The Argument’s literal outrageousness – its existential inaccessibility, its incompatibility with the means we have for arguing it – thus does not prevent it from being true. But what can it mean to say that the Argument is supported by the weight of scientific, experimental evidence? It is here, I think, that we come up against the limits of epistemology. Ray Brassier both grounds and tempers his nihilism by appealing to “our ability to
continually revise our beliefs”, and thereby to accept scientific evidence, even when it is radically counterintuitive. Following and radicalizing Wilfrid Sellars, Brassier says that we must “distinguish the normative realm of subjective rationality from the phenomenological domain of conscious experience”. The latter is altogether delusive, in just the ways that the Argument claims. But the former, according to Brassier, takes place when we “acquire the ability to understand ourselves as rational agents operating in the concept-governed space of reasons”. When we do this, Brassier adds, we find that “reason itself enjoins the destitution of selfhood”. Once we “acknowledge the cognitive authority of… empirical science”, we are also led to reject “the existence of entities called ‘selves’” that have any sort of “autonomous reality”. By separating rational agency from phenomenal selfhood, Brassier claims to avoid the paradox of performative contradiction.

I remain skeptical of Brassier’s rationalistic claim, however. For it is based upon an untenable idealization of the powers of reason. Brassier’s account of “conceptual normativity” is far from the messy ways in which experimental science actually operates. Pragmatically speaking, the Sellars-Brassier normative model fails to encompass the wide range of practices, protocols, instrumental arrangements, and institutions by means of which scientific facts are discovered and established. And conceptually speaking, our rationality only consists – as Bakker suggests in Three Pound Brain – in a set of heuristic devices, which work well enough in their own proper contexts, but which tend to mislead us when they are pushed beyond their limits of applicability. Brassier argues for the possibility, and indeed the necessity, of attaining a “view from nowhere”. But Bakker asks: “How could your ‘view from nowhere’ be anything other than virtual, simply another heuristic?” There is no escaping the double bind of performative contradiction.

Neuropath is devastating and brutal in the way it suggests that science can only impose its harsh truths upon us through an extended, and deeply performative, process. We do not rationally accede to scientific propositions; rather, we are compelled to accept them. We cannot assent to the Argument through any “game of giving and asking for reasons”, as Sellars and Brassier would like to believe. For the Argument is only validated – if that is still the right word – when Neil establishes it in practice, pushing violently against Thomas’ literal inability to accept it.
Brassier argues that, when we are “bound” by a rational scientific claim, “this binding is spontaneously undertaken by a subject, not passively submitted to by an object”. But *Neuropath* narrates a process in which there is no room for any such spontaneity. In other words, we cannot have done with narrative (as Brassier would wish), because only narrative can get us to the point where its own pretensions to establish and stabilize meaning is undone.

We need to *embrace* our performative contradictions, rather than trying to manage or reduce them philosophically. A properly *situated* performative contradiction is crucial, even unavoidable or necessary. Call this a kind of extended or paradoxical pragmatism. The only way to realize the implications and consequences of an idea (which, as William James suggests, are where the idea finds its “truth”) is to enact it. In the present circumstances, since the idea involves our cognitive faculties themselves, it necessitates a kind of experimentation upon ourselves, which results in dislodging ourselves from being in control of the experiment. In consequence, the enactment of the idea necessarily pushes us to the point of performative contradiction. We might say that *Neuropath* sadistically forces its readers to face the consequences of the Argument, in the same way that, within the novel, Neil sadistically forces these consequences upon Thomas. Just as Neil performatively demonstrates the truth of the Argument to Thomas, so *Neuropath* itself performs a symbolic demonstration of this truth for the reader.

How does such an enactment take place within the novel? *Neuropath* continually proposes – or leads us to hypothesize – meaningful explanations for what is going on, and especially for Neil’s violent and bizarre actions. But each time, the book then pulls the rug from under us, so that all of these seemingly plausible explanations end up being discredited. At one point, for instance, Thomas concludes that Neil has “developed, nursed, and concealed some kind of psychopathic affective fixation” upon him. But even as he makes this diagnosis, Thomas also retains the suspicion that it is merely “a way to entangle the unexpected within expectation, to utterly eliminate the threat of surprise”. When he can pigeonhole Neil by assigning a motivation to him, Thomas feels as if “control had returned”. But this kind of assurance never lasts. It is always undercut by some new development, when Neil or another character violates Thomas’ expectations, “moving at right angles to who she was”.

As it repeatedly performs this maneuver, *Neuropath* becomes something of an anti-detective-novel. The detective genre works by moving through a series of particular hypotheses or explanations, each of which is first proposed, and then discredited. The failure of each proposal leaves room for it to be replaced by a better one, until at the end of the novel the case is solved. As Sherlock Holmes says, “when you have eliminated the impossible, whatever remains, however improbable, must be the truth”. But *Neuropath* drags this process into a bad infinity. It never arrives at a determinate explanation, and thereby undermines the very tendency to give an explanatory meaning to what is happening. We end up not knowing what motivates Neil’s actions; nor even whether we should say that he is persecuting Thomas, or rather that he is striving to enlighten him. For we learn, by virtue of the Argument, that such a matter should not be explained in terms of psychological motivations at all.

“There’s no such thing as reasons”, the novel tells us again and again, “just causes”. But it also tells us – as Neil reminds Thomas at one point – that even though “reasons may be deceptions… they’re still functional”. In other words, rather than eliminating narrative altogether, *Neuropath* narrates the failure of narrative expectations. Only narrative can explain how, in Brassier’s words, “the very category of narrative… has been rendered cognitively redundant by modern science”. We cannot get away from reasons and explanations – all the more so when the inevitable failure of reasons and explanations is what we are trying to explain. And similarly, there is no way of getting beyond genre fiction conventions – like those of science fiction or the detective novel – especially when we trying to criticize or undermine them.

Another way to put this is to say that, in its exploration of performative contradiction, the novel strains toward a point that it ultimately cannot reach: a point that cannot be represented or dramatized within it. This point is that of Neil’s own consciousness. What is it like to actually live and experience the world in such a way that Nagel’s “what is it like?” question becomes meaningless? Even as Thomas is strapped to the Marionette machine, we are told, “the greater part of him wondered, even revered. What would it be to walk without self or conscience, even indistinguishable from compulsions, one more accident in the mindless wreck that was the world? What would it be like to act, not as something as puny or wretched as a person, but as a selfless vehicle, a conduit for
everything that came before?” The novel forces us to confront this question, but also prevents us from being able to answer it.

Neil embodies the enigma of performative contradiction, because he claims to have actually attained the state of existing without the illusion of responsible selfhood. “I still experience things, after all”, he says; “it’s just a radically different experience, one far more sensitive to the fragmentary truth of our souls. One without volition, purpose, selfhood, right or wrong”. Neil has apparently “passed beyond the veil… he thinks he’s seen his way through the illusions of consciousness”. In such a state, Neil no longer operates in terms of “motives, goals, reasons”. But rather, as he describes it: “I’ve disconnected certain performance-inhibiting circuits… What you folk, psychologists, call anxiety, fear; all that bullshit. They’re little more than memories to me now. But I’ve also shut down some of the more deceptive circuits as well. I now know, for instance, that I will utterly nothing. I’m no longer fooled into thinking that ‘I’ do anything at all”.

We might like to think of this state as the rational “destitution of selfhood” that Brassier finds crucial to the scientific method, allowing us to operate “in the concept-governed space of reasons”. We might also like to regard it as a kind of Zen selfless attentiveness, such as the consciousness researcher Susan Blackmore describes:

If I genuinely believe that there is no ‘I’ inside, with free will and conscious deliberate choice, then how do I decide what to do? The answer is to have faith in the memetic viewer; to accept that the selection of genes and memes will determine the action and there is no need for an extra ‘me’ to get involved. To live honestly, I must just get out of the way and allow decisions to make themselves. I say the result is unnerving because at first it is odd to observe that actions happen whether or not ‘I’ will them. I used to have two possible routes home, the main road and the prettier but slower lanes. As I drove up to the junction I was often torn by indecisiveness. How could I decide? Which would I enjoy most? Which would be best? One day I suddenly realized that ‘I’ didn’t have to decide. I sat there, paying attention. The lights changed, a foot pressed the pedal, a hand changed gear, and the choice was made. I certainly never went straight on into the stone wall or bang into another car. And whichever way I went was fine. As time went
on I found that more and more decisions were like this. It brought a great sense of freedom to let so many decisions alone.

\textit{Neuropath}, however, doesn’t let us off the hook by offering us any such easy escape hatch. The novel presents a far less benign sense of “I”-lessness than is provided by either Brassier or Blackmore. We learn that Neil has performed surgery both on himself, and on other NSA agents, in order to eliminate all feelings of conscience or responsibility. This is rather hilariously called the “Flat Affect Neuroplasty Program”. Its aim is to turn you into a “radio-surgical psychopath”. Agents have their “amygdalas…stripped down to their predatory essentials”. All their “social circuitry” is “amputated”. They are swept clean of “compassion”, “guilt”, and “shame”. They are now free to operate “without a whisper of self-consciousness”; as far as they are concerned, “anything goes”.

The purpose of the Program, we are told, is “to position flat affect bargainers at every level of the government and military”. The technology works to create “people surgically unaffected by your Stone Age biases. People capable of driving the hard bargains, who don’t need to bullshit themselves when it comes to choosing the projection of US power over the dissolution of the Knesset, or Orinoco drilling rights over starving Venezuelans”. The semantic apocalypse may well be at hand, but “thanks to us, America will survive to pick up the pieces, believe you me”.

The parodic language here lets us know that we are neither in the realm of mysticism, nor in the “space of reasons” – but fully within the strategic logic of the national security state. Selflessness or impersonality is simply another tool for manipulation. The Flat Affect Neuroplasty Program is based on the supposition that, once we eliminate the illusions of selfhood, and once the layers of social conditioning are removed, we will be left with the cold self-interest of a rational “bargainer” who chooses the most efficient game-theoretical strategy. And isn’t this where we must inevitably end up, once we have concluded, along with Daniel Dennett, that “consciousness cannot be separated from function”? Philosophy and science alike are reduced to their use for the machinations of power. It does not matter what consciousness actually is, but only how it can be manipulated.

Yet even this does not seem to go far enough for Neil – which is perhaps why he goes rogue from the NSA. For Neil, game-theoretic calculations are
beside the point. There can be no contest, he says, and no “winners or losers”, because – once the illusion of selfhood has been stripped away – “there’s no one keeping score” any longer. Neil often talks in terms of “hardwired” Darwinian imperatives for survival and the passing on of genes to progeny. But even such bedrock commitments are no longer left in place, once “the rules binding everyday human intercourse” have been swept away as sentimental twaddle. Neil knows that ultimately we are “hardwired” as much for “loyalty and solidarity” as we are for “infidelity” and sexual gratification; not to mention for “self-justifying rationalization” and for being “biased and closed-minded”. The “hardwired” tendencies and traits of the human mind give rise to all our delusions of consciousness – and it is these that Neil’s neurosurgery works to strip away.

In other words, the Flat Affect Neuroplasty Program fails to embrace the full force of the performative contradiction that arises from manipulating and experimenting upon – which is to say from undoing – oneself. Although “the feeling of being you… can be shut down with the flick of a switch”, this is not enough for Neil’s own explorations as “the world’s first neuraonaut”. Pushing things to the limit is no easy matter; it requires a more radical intervention. As Neil tells Thomas, “your brain needs to process the actual loss of its network, it needs to see it crash. Only then will it be able to accept, to see through the cartoon mind it confuses for itself”. As mystics and neuromanipulators both know, there must be a stage of “desolation as insight” before true illumination can be reached. Only trauma can truly liberate me from the “hardwired” illusions of selfhood.

And this is the point towards which Neuropath strives. When the Marionette machine wipes out Thomas’ sense of selfhood, we are told, “there was simply this clearing, this space, a manifold of things and happenings, articulated in time, and belonging to no one”. Within Thomas, “something began to understand. Something… not him”. For now “he was but a moment, something deeper than him realized. Nothing more than a fragment, fooled by blindness into thinking itself whole”. Is this, as Neil claims, “what it’s like when the self is shut down”? Does the first-person “I” still exist, when it is no longer like something to be the way I am? Such oxymoronic language is unavoidable here, for the novel depicts the destitution of selfhood as an infinite approach to an asymptotic point that is never actually reached.
At the end of the novel, Thomas is brought to the verge of this mystical illumination, only to be thrust back to the horrors of everyday life. He cannot help feeling a certain sense of loss at reverting back to himself: “everything would be shadows after this – simulations. No fear, no pain, no joy or love would be as profound, as true, as what Mary had shown him”. Thomas is left like a prisoner in Plato’s cave, or like a cinema spectator, waiting for the final revelatory breakthrough that will never come: “experience, unspooling like a movie, qualms for color, hopes for shape, decisions for the illusion of movement, waiting for the bulb to burn through, for the celluloid to boil into black rings, so that it all could vanish into the hidden frame, leaving only catcalls and white light on a white screen”. Thomas will never experience this nirvana, this nothingness, this access to “the brain beyond and beneath”.

As for Neil, his presumptive mental state shimmers just beyond the pages of the book, something that resists all representation. Or better, we should say that Neil’s condition cannot be figured within the novel, precisely because it is not a something. We should add, though, along with Wittgenstein, that it “is not a Nothing either”; rather, it is that “about which nothing could be said”. Neil’s mental state – if we can still call it that – is the point toward which the entire novel strains, but which it cannot finally inhabit. For such a condition resists being narrated: even though, or perhaps precisely because, it can only be approached and designated in the course of a fictional narration like that of Neuropath. Neil’s brain state, or mental condition, cannot be represented, because it has itself been swept clear of representations. I mean this quite literally: it apparently no longer contains any of the “mental representations” that, according to Metzinger’s orthodox philosophy of mind, constitute both the “content” and the “vehicle” of consciousness. For Neil, the “cartoon” of consciousness – or what Metzinger calls the phenomenal self model (PSM) – “no longer exists”. It might even be the case that Neil is effectively a philosophical zombie. This is a being who acts in all respects (including that of making subjective reports of his inner state) like a conscious human person, but who is actually devoid of any inner experience at all. How could we know?

The enigma of Neil’s inner state – its performative contradiction – cannot be resolved rationally, conceptually, or philosophically. Instead, it is worked out in the terms, and with the resources, of genre fiction. That is to say, since it cannot be meaningfully integrated into the narrative, it has to be
killed off, or wiped from the narrative, instead. *Neuropath* ends, as it must, with Neil’s death. He is murdered by his own initial victim, Theorodos Gyges. Through a necessary and profound irony, Gyges is able to find Neil, and take revenge upon him, despite not being able to recognize him. It is only at this point that we finally find out that Gyges does not only suffer from face blindness, but is also the sexual serial killer known as the Chiropractor.

This ending is not just an arbitrary *deus ex machina*, because Gyges’ own backstory turns out to be crucial. We learn that Gyges is caught in the contradiction between the rape-murders that (thanks to Neil’s surgery) he compulsively performs, and the religious sensibility that he still retains. He has no access to the *causes* of his actions, but he nonetheless still needs to give himself *reasons* for what he does, in order to excuse his own behavior. To justify his actions to himself, The Chiropractor always begins his assaults by severing the victim’s spinal cord. In this way, he says, although “they know” what is happening to them, “they do not *feel*” it – or, in the revised version of the novel, “they can’t *feeeeeeel* what I do. That means it’s *not* a sin, doesn’t it? *Doesn’t it?*” Gyges wants to believe that he isn’t really hurting his victims, because he takes care that they do not suffer any physical pain. When the spinal cord is cut, he says, “the soul is preserved, kept safe, wrapped in a box”. Gyges proclaims that “I only fuck the *meat*”; and the meat, for him, is finally not of any importance.

Just as Neil intended, Gyges embodies the Argument, or *lives* it. Even though he is trapped in an endless nightmare in which he cannot recognize himself, and even though he is continually compelled to perform monstrous actions that he does not countenance, Gyges nonetheless produces a fabulation that makes it all seem meaningful and reasonable. This is echoed in the way that his victims are still forced to *know* the horrors that they are unable to *feel*. Neil performs the inverse operation when he subjects Thomas to the Marionette machine. For Thomas is compelled to *feel* his own inability to know. The Argument demolishes all our pretensions to self-understanding, or indeed to any sort of positive knowledge. Cognition is nothing more than delusional confabulation, or rationalization after the fact. But we cling so desperately to our “cartoon” illusions that cognition can only break down when we are made to feel the full force of its implosion. Thomas is swamped by overwhelming waves of impersonal affect:
Never had he so yearned, as though a chasm had cracked open within him, an endless clutching abyss, suddenly filled with divinity, with a resounding, weeping unity, pinged by twinges of anxiety that grew like bloodstains, that blackened into a thrumming dread, with claws like capillaries, peeling muscle from the inside of skin, while the world before him flapped back and forth like wings on an interdimensional hinge…

This is the point at which the scientific reduction of consciousness flips over into a kind of ferocious mysticism, a *via negativa*, a mortification of both the mind and the flesh. Of course, this experience is an entirely instrumentalized one. It corresponds to no beyond, and it can be turned on or off with the flip of a switch. But even in this nullity, this absence of self—where there is “no feeling, no sensation”—there is still an incipience: “just a trembling, a teetering blacker than black”.
Michael Swanwick’s short story “Wild Minds” (1998) reads like a riposte to Scott Bakker’s *Neuropath* – even though it was written a decade earlier. The story strives to find continuing value in the adventure of human consciousness, even if all the darkest suspicions raised by reductionist neuroscience and eliminativist philosophy of mind turn out to be true. Swanwick insists upon the worst. He posits a world in which “the dark dimensions of the human mind” have been purged and exterminated, and he strives to affirm them nonetheless.

The story is set at a near-future point in which “the workings of the human brain” have been “finally and completely understood” by science. The story presents this as a *fait accompli*. Everything psychological, people now understand, is simply a matter of maintaining the right “chemical balances”. Everyone learns “the structural basis of emotions, and how to master them before they flush the body with adrenaline”. Everyone also knows that “self is an illusion. The single unified ego you mistake for your ‘self’ is just a fairy tale that your assemblers, sorters, and functional transients tell one another”.

In such a world, there is no longer any need for the old common-sense views of the mind that the eliminativist philosophers condescendingly dismiss as *folk psychology*. According to these philosophers, there are no such things as *beliefs, desires, and feelings*. These are just misapprehensions, or after-the-fact fabulations. When people claim that they are in pain, for instance, the *real* fact of the matter is that their C-fibers are being stimulated, and that they find this stimulation unpleasant. More generally, as Paul Churchland puts it:

> Our common-sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will
eventually be displaced, rather than smoothly reduced, by completed neuroscience.

For Churchland, as for Bakker, our introspective reports on our own mental states are entirely delusional, and radically false. An accurate scientific picture of the mind will not subsume our habitual picture of ourselves, in the way that modern physics subsumes Newtonian physics as a special case. Rather, it will force us to entirely reject “folk psychology”, and construct our understanding of the brain on a totally new basis.

In the world of the story, this new basis has been definitively established, and is accepted by everyone. The people in the story describe their moods, for instance, as processes in which “some emotional sea-change” is “organizing itself deep on the unseen levels – the planners building new concept-language, the shunts and blocks being rearranged”. They apologize for giving expression to “emotional transients” that are not being kept sufficiently “under control”. Instead of providing fabulated reasons for their own – and other people’s actions, they now cite the actual physical causes that produce those actions. People say things like “my assemblers and sorters got into a hierarchic conflict”. They don’t fall in love so much as they discover that their “emotional components” are threatening to “collapse into a new paradigmatic state”.

It’s not only self-perception that changes in the wake of the revolution in neuroscience, but all sorts of other things as well. For instance, traditional forms of education are no longer necessary. “Mere learning” is far too “easy” to make a fuss over. Any conceivable knowledge or skill can be implanted quickly, through direct electrochemical stimulation of the brain. “Anybody could become a doctor, a lawyer, a physicist, provided they could spare the month it took to absorb the technical skills”. There is no longer any notion of learning “for its own sake”; nor is formal education a job requirement. Instead, “most corporations simply educated their workforce themselves to whatever standards were currently needed”.

You might think that such a development would at least equalize people’s social and economic opportunities. After all, nobody would be left behind any longer, as a result of lack of talent, inadequate training, or childhood deprivation. Anyone can learn to perform the same tasks just as well as anyone else. The complete and transparent understanding of the brain should lead – in principle at least – to the unleashing, and full
realization, of what the Italian autonomists call general intellect: widespread social knowledge, or collective intelligence, which, in the post-Fordist era, replaces labor as the “main productive force”. It should also lead to the final realization of Jacques Rancière’s axiom of equality, which states that “the same intelligence is at work in all the acts of the human mind”.

But in Swanwick’s decidedly non-utopian story, things do not turn out this way. The new science is not liberating, and it does not lead to any greater measure of social equality. Rather, it works to reinforce hierarchy, and to intensify corporate control. This is a pattern that’s all too familiar today. New technologies – like our radical advances in computation and communication over the past few decades – do, in fact, offer us new potentialities for liberation. But their actual effect has largely been to increase the power of corporations and governments. The same is true of the current “second machine age”, with its promise of automating jobs, from driving trucks to customer service to generating mathematical proofs. These inventions could well lead to a world of universal affluence and light work schedules. But they will most likely be used, instead, to increase the wealth of a small number, while more and more people are unemployed and impoverished.

The same is true in the world of “Wild Minds”. With access to the tools of a completed neuroscience, businesses are all the more ruthlessly selective about whom they will hire, and under what conditions. Since they no longer need to worry about their employees’ lack of ability, they have a freer hand than ever before. The lowered cost of knowledge, skill, and training – like the cheapening of other factors of production – leads to new opportunities for squeezing out profit. Marx argues that the labor-cheapening process under capitalism is limited by what he calls the tendential fall in the rate of profit: as productivity increases, and formerly scarce goods become cheap and abundant, prices tend downwards – towards zero – and profit margins become ever smaller. But this tendency can be offset, as Marx notes, by a number of “counteracting factors”. Indeed, modern capitalist enterprises have found all sorts of ways to defer any such outcome indefinitely. Most notably, they have succeeded in re-introducing scarcity and austerity in the face of their own production of abundance.

And so, in the world of “Wild Minds”, competition among workers becomes harsher than ever. “With knowledge so cheap, the only thing
workers had to sell was their character: their integrity, prudence, willingness to work, and hardheaded lack of sentiment”. Since employment is no longer a matter of differing capabilities, it becomes all the more a matter of proper character. Employees must possess those particular personality traits required by what Luc Boltanski and Eve Chiapello call “the new spirit of capitalism”. Workers need to be enthusiastic, flexible, amenable to orders, and willing to “take responsibility”, all at once. Specifically, the workers’ “willingness to work” must be so great that they don’t leave anything behind at the office, but devote themselves 24/7 to their assigned tasks. Their having “integrity” and “prudence” means that they will never put their own interests – let alone those of a group, or of anyone else – ahead of the job and the corporation. At the same time, they must be sufficiently “hard-headed” that they never question their orders. And their “lack of sentiment” means that they fully accept the precarity of their employment. They must also be able, once a given project has been completed, to entirely shift their loyalties, so that they can easily move on to a new project with the same enthusiasm as they had for the previous one, and without feeling any twinges of regret or nostalgia.

In “Wild Minds”, therefore, the completion of brain science makes it easy to produce people whose minds are carefully cultivated instead of remaining “wild”. You can easily be implanted with the very character traits that corporations demand:

It was discovered that a dozen spiderweb-thin wires and a neural mediator the size of a pinhead would make anybody as disciplined and thrifty as they desired. Fifty cents worth of materials and an hour on the operating table would render anybody eminently employable.

This process is called optimization. It takes place through a simple “outpatient operation… no more complicated than getting your kidneys regrown”. Of course, we cannot actually regrow kidneys today; but presumably organ regeneration is a simple matter in this future world. The progress of medicine over the past century has been astonishing; we have no reason to doubt that this progress will continue, or that the methods that work so well for fixing hearts, kidneys, and other organs will work for manipulating brains as well.
In the world of the story, even unoptimized people know that “folk psychology” is a false, discredited theory, and that their intuitions regarding their own minds are nothing more than what Scott Bakker calls “artifact[s] of medial neglect”: the inability of consciousness to track the very processes that generate it. But the optimized no longer experience these “artifacts” at all; optimization actually eliminates the illusions of selfhood that generated folk psychology in the first place.

That is to say, optimization gives you complete insight into your own mental processes. In Bakker’s novel, this is the extreme state apparently attained by Neil Cassidy. In Swanwick’s story, such insight is much more easily obtained. Once you have been optimized, you have “absolute clarity of thought” – indeed, to such an extent that you become “totally unable to lie to [your]self”. You are actually able to engage in “rational discourse”, in sharp contrast to the unoptimized “wild minds”, who only “care about winning” an argument, and who suffer from confirmation bias. With optimization, you also gain “control over involuntary functions”, together with “freedom from prejudice and superstition”, and “freedom from the tyranny of emotion”. In addition, you have “the ability to block pain”, and “accelerated regenerative ability”, to repair the damage that pain signals – or rather, stimulated C-fibers – originally evolved in order to warn us about. And you do not ever need to sleep. All this is clearly a boon for increased economic productivity.

Optimization thus makes for perfect corporate employees. There are no conflicting loyalties to worry about. Not only does optimization put you in direct control of your emotions; it also leads to the dissolution of any previous ties of sympathy and concern. Religious belief is “the first thing to go”; the delusive and compensatory quality of religion becomes too glaringly obvious. The optimized mind no longer has any need for it. Family ties, and other sorts of empathy, come next; and aesthetic passions quickly follow. They are all rejected by the illusion-freed mind. As the narrator tells it, when his wife was optimized:

She was still working things out when she came home [at 6 pm]. By seven, she’d seen through God, prayer, and the Catholic Church. By eight, she had discarded her plans to have children, as well as a lifelong love of music. By nine, she’d outgrown me.
With his portrayal of optimized minds, Swanwick seems to be parodying – in advance! – the ideals of today’s so-called New Atheists, like Richard Dawkins, Sam Harris, and the late Christopher Hitchens. These thinkers love to describe themselves as “brights”. This seems to mean that they regard themselves as being entirely rational, logical, and free of illusions – in sharp contrast to the ostensibly benighted masses. Their attacks on religion are so crude and grandiose as to suggest that they have no understanding whatsoever of the affective bases of religious belief – or of any other characteristics of “folk psychology”, for that matter. They brutally dismiss what they confidently know to be wrong. The New Atheists’ self-congratulatory smugness is a quality shared by optimized people within the story as well. When the narrator asks his optimized acquaintance to come to Mass with him, she “looked at me as if I’d invited her to wallow in feces. Then she laughed. ‘Will I have to eat human flesh?’”

A remark like this speaks for itself, even if the speaker later withdraws it, asking forgiveness for having been “unspeakably rude”.

Even though optimization is a simple procedure, available in principle to everyone, only a small number of people actually choose to go through with it. “The Bureau des Normalisations et Habitudes”, we are told, is “afraid that not enough people [a]re signing up for optimization”. This goes back to the time when optimization was first invented:

The ambitious latched onto optimization as if it were a kite string that could snatch them right up into the sky. Which, in practical terms, it was. Acquiring a neural mediator was as good as a Harvard degree used to be. And – because it was new, and most people were afraid of it – optimization created a new elite.

There are clear class differences between the people who get themselves optimized, and those who don’t. Ordinary, unoptimized people are the proles, the working class. You don’t get very far when you are unoptimized, because “you’re constantly at the mercy of forces you don’t fully understand”, brain states that you know to be mere physical contingencies, but whose actual, causal details remain obscure to you. The unoptimized people are losers in the great neoliberal race of economic competition. Like non-college graduates today, they usually cannot get good jobs. And they also tend to live in places like Glasgow, the narrator’s home, a city whose
“logic is essentially medieval: The streets have grown as they will, in a rough sort of grid, and narrow enough that most are now fit for one-way traffic only”. In short, the neighborhoods of the unoptimized are run down, uncomfortable, and inconvenient. They have never been redeveloped:

Old places… cluttered with seedy pubs and street corner hangouts, the niches where shabby men sit slumped over their whisky in paper bags, the balconies from which old women watch over the street.

The optimized “new people”, on the other hand, are “unnerv[e]d” by “this stench of accommodation and human dirt”. They do everything they can to avoid such “primitive squalor”, and to keep themselves as separate as possible from the “obsolete people” who have not had the operation. Indeed, the optimized are virtually a new species; they “don’t claim to be human” any longer. They have gone down the rabbit hole, embraced the Singularity, and passed beyond the point of no return. They are nonetheless confident that, despite this breach, the remaining all-too-human society cannot possibly do without them: “do you have any idea how complicated the world has gotten? Unaugmented minds couldn’t begin to run it”.

Like yuppies and entrepreneurs today, the story’s optimized people value immaculately clean streets, reflective surfaces, and sleek, modernist architecture. Everything in their world is “bright and fast”, so perfectly articulated that even “the air sings”. They live in buildings with “shimmering planes and uncertain surfaces… buildings that could never have been designed without mental optimization, all tensengricity and interactive film”. The word tensengricity does not seem to be in the dictionary; I assume it is a version of tensegrity, a word invented by Buckminster Fuller to describe an architecture in which structures are stabilized by means of carefully placed materials in strong tension. It would seem that the optimized people demand dynamism and stability, without wastefulness – all at the same time. They also expect “interactive” environments that fully and unequivocally respond to their whims.

Thom, the narrator of “Wild Minds”, is an outlier in this society. He is evidently a businessman of some sort, a “salesman” with a complex “travel schedule”. And yet, despite his affluence and class standing, he “won’t accept optimization”. His resistance is the main point of the story. In “Wild Minds”, Thom recounts the night he spends with Hellene, a “human
resources” person and “corporate recruiter”. She is a proselytizer for the new technology. She tries to convince Thom to get optimized, but he refuses. At the end, she departs. Very little actually happens in the course of the story; but Thom reflects on his own past, and tells us what has brought him to such a point.

Thom meets Hellene at “a businesspersons’ orgy in London”. This is itself a telling detail, even though the orgy is only described briefly, and without any hardcore details. The event is apparently a standard part of any business executive’s schedule. It is carefully monitored, to make sure that it fits the requirements of psychological health. Thom notes that the robotic “doorkeeper” would not let him in until “it saw that I wasn’t acting out a sex-addiction script, but properly maintaining my forebrain and hindbrain balances”. The orgy thus works as a therapeutic release for everyone involved. It allows them to indulge their sexual desires, without the need for demanding emotional commitments. Unsurprisingly, then, this orgy is not frenzied, but entirely polite and respectable, and fairly low-key: “the light was dimly textured and occasionally mirrored. Friendly hands helped me off with my clothing”. People “quietly” give their names; Thom just notes that “time passed”.

It is also telling that Thom first becomes aware of Hellene “not because she was beautiful – who pays attention to beauty, after the first hour? – but because it took her so long to find release”. This seems oddly dysfunctional. Thom realizes at once that Hellene is optimized. He also notes that “she had those chill Scandinavian features that don’t show emotions well”. But Hellene explains to him that she is finding it hard to get off, because her “neural mediator has become unreliable”, so that she has been forced to take it “offline”. She must also wait for several days before she can get an “upgrade”. In the meantime, she cannot stop herself from losing control and getting all emotional. She finds this highly embarrassing; she is “not used to functioning without the mediator”.

Hellene is also worried – all the more so when her mediator is not working – by the chore of having to visit her children. She is not looking forward to this, she says, and neither are they. When people are optimized, they tend to reject having children in the same way that they reject religious belief. Once your illusions have been dissipated, and you can clearly see what efforts and sacrifices will be required, the idea of having offspring loses its appeal. The only way that optimized people will agree to have
children is by offering them what the neoclassical economists like to call incentives:

It was discovered that optimized people weren’t having children, so they crafted a regulation giving serious career preference to those who did.

In consequence, Hellene decides to have children even though she doesn’t particularly want them. She is being recompensed for her annoyance; it’s a simple matter of economic calculation. It would seem that optimized people actually function according to the dictates of rational choice theory – even though unoptimized human beings do not. As for the children themselves, Hellene “signed them up with Sterling International for full optimization when they were eight”, to make sure they would have the right status and the right opportunities. And – more incentives! – she also tells Thom that

I face a severe fine if I don’t see them at least twice a month. It’s happened three times so far this year, and quite frankly, my bank account can’t take it.

On the basis of all this, Thom considers Hellene “a bad mother”. But she is unperturbed by this judgment; she takes it as just another sign of how old-fashioned Thom is.

Hellene cannot for the life of her understand why Thom has not been optimized. “You seem an intelligent enough man”, she says to him; “why reject what science has revealed about the workings of the brain?” Thom tells her that the Catholic Church “considers [optimization] a mortal sin”, in a way that merely attending an orgy is not. You can continue to do the latter, “as long as you go to confession before you take Communion”. But optimization remains beyond the pale. Thom also tells Hellene that he is “afraid of losing myself” – even though he concedes that the very notion of “self” is an illusion to begin with. Hellene is unconvinced by these objections. As she says to him, “I don’t like ambiguity. It’s an artifact of the old world”. On the other side, Thom is forced to conclude that, for all of Hellene’s optimized insight, “she just didn’t get it”.

Hellene tries hard to convince Thom to be optimized. She gives him what he recognizes as her standard recruitment speech, sensing the
“smooth, practiced quality to her words”. Finally, she offers him a “prototype recruitment device”. By means of a “magnetic resonance simulacrum”, it will show him, “for the space of fifteen seconds… how it feels to be optimized”: what it is like from the inside, in other words. She promises that, after the fifteen seconds, he will revert to his previous mental condition. Hellene feels certain that this demonstration by experience will convince Thom that “there’s nothing to be afraid of”.

Thom agrees to take up Hellene on her offer. But he asks her, in return, to accompany him to Sunday morning Mass. If he is going to briefly experience her mode of being, he wants her similarly to get a sense of the religious experience that she has rejected, and that he clings to. He has already told Hellene that he meditates. She characteristically responds that “you wouldn’t need meditation if you were optimized”. The peace of mind he seeks would just come to him automatically. Hellene asks Thom, “What do you see when you meditate?” He replies: “Sometimes I see comfort there; other times I see suffering”. Optimization would presumably resolve the ambiguity by entirely removing the latter. But of course, that is precisely the problem with optimization, from Thom’s point of view.

Thom uses, as his “focus” for meditation, a replica of a painting by Ad Reinhardt. Thom’s copy of the painting “is a duplicate as exact as human technology can make it; more exact than human perception can distinguish”. The painting is called “For T. M.”; it is an abstract expressionist work from 1957, one of Reinhardt’s “black paintings”:

At first it seems unvaryingly colorless; you have to stare at it for some time to see the subtle differences in the black, the thick cross that quarters and dominates that small lightless universe. He painted it for Thomas Merton, who was a monk.

The painting is a severe and subtractive work. Reinhardt seeks to remove as much as he possibly can from the painting. There is still a figure-ground relation, but it is barely noticeable. Other common qualities of visual images – like texture, color, and light – are entirely excluded.

Thom is also proud to have a Charles Rennie Mackintosh chair. He calls it “an original because it was made to his directions”. In other words, the chair is also a replica. But for the chair as for the painting, any difference from the actual original is too subtle for human perception to distinguish.
Thom places the chair opposite the painting. “Sometimes I’ll sit in the one and stare at the other”, he remarks, “thinking about distinctions, authenticity, and duplicity”.

The reductionist severity of the Reinhardt painting, and the decorative Art Nouveau sensibility of the Mackintosh chair, are equally opposed to the “bright and fast” aesthetic of the new, optimized people. The painting is, if anything, as extreme and reductive as the optimization process; but its brooding negativity moves in the opposite direction, offering murky “depths” instead of vibrant surfaces. The “straight-back” chair is built to human scale, in proportions that the optimized people have moved far beyond.

Thom isn’t entirely sure about his own values, as they are manifested in these works of art. If you haven’t been optimized, you cannot be free from doubt. Thom knows that his retro-modernist aesthetic means that he is, in fact, clinging to his doubt, rather than seeking to resolve it. Is there any way around this? At one point in the course of the evening, Thom says, “I felt like a feral child standing on the twilight lands between the cultivated fields and the wolf-haunted forests, unable to choose between them”. He is tempted by Hellene’s arguments, even though he rejects them. And he is tempted by Hellene herself, perhaps because “she reminded me of Sophia”, his former wife.

By putting on the simulation device that Hellene offers him, Thom tries out the very perspective that he fears. Optimization couldn’t be more opposed to Thom’s own previous taste, values, and predilections. But he accepts Hellene’s insistence that he cannot really know what he is rejecting, unless he tries it out at least briefly. Mental clarity can be simulated as fully and powerfully as paintings and chairs.

The result of using Hellene’s device is very nearly miraculous. Thom tells us that

it was as if I had shrugged off an enormous burden. I felt myself straighten. My pulse strengthened and I breathed in deep, savoring the smells of my apartment; they were a symphony of minor and major keys, information that a second ago I had ignored or repressed…

The experience was wonderful. Like standing upon a mountain top facing into a thin, chill wind. Like diving naked into an ice-cold lake
at dawn. I closed my eyes and savored the blessed clarity that filled my being.
For the first time in as long as I could remember, I felt just fine.

And that is itself the problem. Thom does not want to feel fine; he believes that he does not deserve it. This is precisely why he hates optimization. It’s at this point that he tells us his backstory, explaining the reasons for (rather than the causes of) his rejection of the technology. It turns out that Thom murdered his wife. They had a long history of arguing: “we yelled at each other for hour upon hour, evening after evening. Sometimes we broke things”. But the tipping point came when Sophia got herself optimized, despite Thom’s objections; in disgust and anger, he killed her.

Thom tells us nothing more than this about the circumstances of his act. But his sense of guilt is the bedrock of his character. So he is appalled by the feelings of lightness and peace that optimization would offer him. And he is convinced that Sophia, too, would “never have chosen optimization knowing it would be like that”. If she had known, she wouldn’t have gone through with it; and if she hadn’t gone through with it, he would never have killed her, and she “would be here with me now”. The problem is, that once you have gone through with it, you are no longer able to revert to what you were before. Once you have achieved the optimized state, you are no longer the person who would have hated and rejected it.

When you’re optimized, you realize that (just as Nietzsche says, and as The Argument in Bakker’s Neuropath states) there is no such thing as “free will”, and hence no responsibility. You don’t just think this abstractly; you know it to be the case. For the fifteen seconds of the simulation, Thom finds himself not regretting a thing. I knew it wasn’t my fault. Nothing was my fault, and if it had been that wouldn’t have bothered me either. If I’d been told that the entire human race would be killed five seconds after I died a natural death, I would’ve found it vaguely interesting, like something you see on a nature program. But it wouldn’t have troubled me.

The complete understanding of the brain also renders traditional notions of guilt, crime, and punishment irrelevant – as well as any idea of agency or
responsibility. Thom also recalls that, after the murder,

a panel of neuroanalysts had found me innocent by virtue of a faulty transition function and, after minor chemical adjustments and a two-day course on anger control techniques, had released me onto the street without prejudice.

This rational understanding of his crime leaves Thom without hope, and in desperate need of “the consolation of religion”. He embraces his Catholic heritage, to which he had previously only given lip service, precisely because “according to the Church, I had sinned. I had sinned, and therefore I must repent, confess, and atone”. Having “performed an act of true contrition”, he now believes that “God has forgiven me”. But he remains unable to forgive himself. And that is “why I’ll never be optimized”, he says. “The thought that a silicon-doped biochip could make me accept Sophia’s death as an unfortunate accident of neurochemistry and nothing more, turns my stomach”. Not only would optimization place him in a situation that he abhors; it would also make it impossible for him to feel that abhorrence any longer.

And so, Thom is a conscientious objector to the new brainoptimized society. Or, what is perhaps the same thing, he is an ironist. He would rather live with ambiguity than reject it. He accepts that “being human” is no longer “essential”; yet he “cling[s] to the human condition anyway, out of nostalgia perhaps but also, possibly, because it contains something of genuine value”.

Is this anything more than an old-fashioned, retrogressive humanism? “Wild Minds” works as a story because its defense of the “human” against the “posthuman” is so nuanced and hinged with irony. The story wouldn’t be in the least convincing if it just preached the eternal verities of the human condition in the usual pompous and high-minded terms. Thom solicits our sympathy because of his having murdered Sophia, rather than in spite of this. The source of his objection is not that there is some sort of flaw in the optimization process, but rather that there isn’t.

Swanwick makes the traditional humanist argument as difficult as possible, for himself as well as for us. Though Thom worries about his own “nostalgia”, he never challenges the premise that science and technology are in fact capable of grasping and mobilizing all there is to know about the
human mind. No appeal to our supposed essential inner being and spirit is able to undermine – or even temporarily interrupt – the scientific discovery, at an ever-accelerating pace, of the actual ways that our brains work. There is no getting beyond the material (biochemical or neurological) basis of thought. In this way, “Wild Minds” responds (in advance) to Bakker’s challenge by conceding all the implications of his deflationary account of human consciousness. The story supports a humanism that is asserted, not in spite of, but precisely through, all the deficits that reductionist neuroscience and eliminativist philosophy of mind are able to discover.

In other words, “Wild Minds” never makes the claim that the optimization that it describes is impossible, or that it fails to deliver on what it promises. If the process described in the story leaves aesthetics and religion behind, this is because these modes of experience are intrinsically delusional and deceptive, and therefore not susceptible to optimization. They remain singular and irrational. They cannot be “smoothly reduced” to neuroscience, but rather will have to be “displaced” and entirely abandoned. “Wild Minds” sides with what has been – or is going to be – discredited and abandoned. If the old manifest image of humanity nonetheless “contains something of genuine value”, this is precisely because “value” per se has no place in the new, mentally optimized world – but only in the “obsolete” one.

“Wild Minds” thus leaves us with the insight that posthumanity – as so many of us have imagined it over the past several decades – is largely a corporate fantasy of control. The posthuman project envisions re-engineering “human nature” in line with the demands of neoliberal capitalism. This means that the discoveries of neuroscience – as they are actually being made today, as well as how they are extrapolated in the story – are pragmatic and operational ones, rather than being essential or foundational. For there is no such essence. Neuroscience is oriented towards power and efficacy, towards the ways that the brain can be transformed, manipulated, and controlled. In refusing optimization, Thom acknowledges (far from questioning) the power of such a procedure. And thereby, he challenges us to imagine – even if he himself cannot – a posthuman transformation that would not merely serve the agendas of capital, or of what used to be called “instrumental reason”.
SIX

Thinking Like an Alien

“By the time I get home, I could be the only sentient being in the universe”. So says Siri Keeton, the narrator of Peter Watts’ novel *Blindsight* (2006). Siri is not being facetious. He is all alone in a tiny spacecraft, making his way back to Earth from the Oort cloud. The voyage will take decades. Siri is the sole survivor of a “disastrous encounter” with an intelligent and technologically advanced alien species from another star system. The human explorers call them *scramblers*. The spaceship *Theseus* is sent out from Earth to investigate the scramblers, and make contact with them if possible. Everything turns upon how the aliens think, what they want, and whether we can find a way to communicate with them. Is there any common ground that we can share? By the end of the novel, all human efforts to contact the aliens, or even to make sense of them, have failed. It turns out that, even though the scramblers are superior to human beings both in strategic skill and in technology, they are not *conscious* in any way that we can understand. Hostilities eventually break out, and *Theseus* and the alien vessel *Rorschach* are both destroyed. Siri’s retrospective mission is to “bear witness” to this cataclysm, to inform the rest of humanity about the aliens and their threat. People back on Earth “don’t know how the story ended”, Siri says. “The lack of closure must be driving them crazy”.

Or maybe not. For something strange seems to have happened back on Earth – judging by the radio communications that Siri is able to monitor from his spacecraft. There used to be “a million overlapping voices” saturating the airwaves, engaged in banter, gossip, and argument. But now all that Siri hears is “traffic control and telemetry”. What can it mean that idle chatter has given way to bare data transmission? Perhaps human beings are not invested in narrative any longer; perhaps we have finally outgrown what Ray Brassier scornfully describes as “our psychological need for stories that unfold from beginning, through crisis, to ultimate resolution”. Perhaps people have finally dispensed with what the philosophers call “folk psychology”. Perhaps we no longer have any use for what Wilfrid Sellars
calls “the manifest image of man-in-the-world”. In any case, Siri is afraid that there may be nobody left on Earth “to take an interest” in his account: nobody caught up in the suspense, nobody even caring how it all turns out. “Maybe the whole exercise is pointless from the start”, he thinks; “maybe no one’s even listening”. For now, Siri recognizes that his story – the words that he is broadcasting, the novel that we are reading – is just “a tale told to myself”; or better, “a memoir told from meat to machinery”, words dictated for digital storage and transmission.

Blindsight is a First Contact narrative, set in the late 21st century. What human beings come to call the Firefall takes place at “1035 Greenwich Mean Time, February 13, 2082”. The Earth is surveyed by a “fleet of 65,536 probes evenly dispersed along a lat-long grid that barely left any square meter of planetary surface unexposed”. (This number, of course, is 2 to the 16th power; the use of binary implies a logic of computational precision.) The probes all burn up simultaneously, after sending their data out into deep space. For the first time, human beings are forced to recognize that we are not the only intelligent, technologically-advanced species in the cosmos. Theseus is dispatched to the far reaches of the solar system, in order to find out who sent the probes.

In the year 2082, when the novel starts, the people of Earth are blessed with a “shiny new post-scarcity economy”. Nonetheless, the world is still controlled by large corporations, together with state military and security services. The highest values still seem to be optimization and streamlined efficiency. Scientists have even spent years, unsuccessfully, in trying to rid human beings of the need for sleep:

The waste was nothing short of obscene: a third of every Human life spent with its strings cut, insensate, the body burning fuel but not producing. Think of all we could accomplish if we didn’t have to lapse into unconsciousness every fifteen hours or so, if our minds could stay awake and alert from the moment of infancy to that final curtain call a hundred twenty years later.

Since they cannot be optimized beyond a certain point, human beings are increasingly relegated to the status of disposable commodities. “Human nature was becoming an assembly-line edit”, Siri tells us; “Humanity itself increasingly relegated from Production to product”. Since digital machines
perform most tasks better than we ever could, human beings have become “redundant in unprecedented numbers”. There are few jobs available; there is little for anyone to do. The result, for most people, is boredom and an overwhelming sense of futility. One character sardonically describes herself as “a parasite on the Body Economic”.

One way that people seek to alleviate their lack of prospects is through posthuman augmentation. They tweak their bodies and extend their senses with all sorts of “retrofits and enhancements”. These fixes involve genetic alterations, neurochemical tweaking, and the implantation of microchips and other digital machinery into the body. Such radical reengineering equips people for specialized technical tasks: those rare activities that still cannot be performed with sufficient accuracy by machines alone. As Siri says of himself and his Theseus crewmates, “the only reason we were here was because nobody had yet optimized software for First Contact”.

In the world of the novel, technology no longer works (if it ever did) in the manner described by Bernard Stiegler: prosthetically, to extend human capacities, or to compensate for an originary human deficit. In the world of Blindsight, the situation is just the reverse: human bodies and minds themselves only work to extend the power and intelligence of machines. In the prophetic words of Marshall McLuhan: “Man becomes, as it were, the sex organs of the machine world, as the bee of the plant world, enabling it to fecundate and evolve ever new forms”. By retooling themselves to perform highly advanced and delicate duties, as adjuncts to digital devices, a few people can still “attach some pretense of usefulness to their lives”. An existence fecundating machines is not very fulfilling; but it is a “lesser evil”, Siri suggests, than simply living “the life of a parasite”.

Siri describes Theseus as “a boatload of freaks”: its crew members have all moved so radically beyond the biological “baseline” as to scarcely be human any longer. In Blindsight and its sequel Echopraxia, Watts gives us a catalogue of potential posthuman augmentations, together with pathological conditions that disrupt and distort human mental functioning. The line between pathological dysfunction and superhuman ability is not always easy to draw; one may also entail the other. Indeed, the very augmentations that give posthuman bodies their extraordinary powers may well also leave them incapable of the simplest everyday tasks.

Consider, for instance, the biologists aboard Theseus, Isaac Szpindel and Robert Cunningham. They have both turned themselves into cyborg
mechanisms, their “consciousness… haunting remote sensors”. They no
longer have to interpret readouts from digital devices; instead, they can feel
the data directly, “right down in the bone”. In this way, they extend their
awareness far beyond the regular human sensorium. Szpindel and
Cunningham have each “butchered one body” – their own – “to become a
fleeting tenant of many”. Szpindel has been “spliced and diced, a gangly
mass of tics and jitters that could barely feel his own skin”; and
Cunningham is

so massively interfaced with machinery that his own motor skills had
degraded for want of proper care and feeding; he heard x-rays and
saw in shades of ultrasound, so corrupted by retrofits he could no
longer even feel his own fingertips without assistance.

Siri has difficulty regarding Szpindel and Cunningham as complete human
beings, because so many “pieces of [them]” are “hidden in machinery”.

In Blindsight, we also encounter straight-out mental aberrations, which can
be induced by lesions to the brain, but also – as happens in the course of the
novel – by exposure to powerful magnetic fields and radiation. One of these
aberrations is the eponymous condition of blindsight, in which someone
who is cortically blind is nonetheless able to respond to visual cues that
remain outside of conscious awareness:

Something in your head is still taking it all in. Something in the brain
is still seeing, and hearing, even if you’re not – aware of it.

Another is Cotard’s Syndrome, in which someone is delusionally convinced
that they are dead. Yet another is inattentional blindness, in which people
cannot see what is immediately in front of them, because their attention has
been directed elsewhere.

Watts mentions in his “Notes and References” to the novel that “many of
the syndromes and maladies dropped into Blindsight I first encountered in
[Thomas] Metzinger’s book” Being No One. This makes perfect sense. For
as Graham Harman suggests in his discussion of Metzinger,
by showing how much complexity is underway in our supposedly simple selves, Metzinger leads us to conclude not ‘well then, the self is just a sham in the end’, but ‘think of how many different and bizarre selves we might create, or which might already exist among animals or on other planets!’

In short, rather than turning the self into a fictional unreality he turns it into a science fictional reality, in which the human is just another bizarre species whose experience is generated by specific constraints, just as reptiles, insects, and extraterrestrials might have different lives from ours at this very moment… Metzinger peoples the possible universe with a vast number of phenomenal species, and even borders on a science fictional version of what I have called ‘speculative psychology’.

Blindsight is Metzingerian speculative fiction in just this way. Metzinger himself concludes from such delusional cases that our “naive folk-psychological notion of ‘consciousness’” must be wrong. Our ordinary mental states are every bit as mediated and constructed as these extreme conditions are. In ordinary everyday life as much as in pathological situations, we are blinkered by what Metzinger calls “the illusion of naive realism: the inability to recognize a self-generated representation as a representation”.

However, the fact that these bizarre mental states are objectively delusional does not prevent them from actually being felt and experienced. And this is what we get in Blindsight. The novel explores the phenomenal feel – or the what-is-it-likeness – of just such bizarre mental delusions. We cannot actually “know” such syndromes directly: for to be caught in one of them is precisely to be unable to grasp it as a syndrome or a mental construction. Indeed, these extreme conditions most likely cannot be contained within our usual understanding of what it means to be “human”. The novel concretely exemplifies David Roden’s disconnection thesis: posthuman entities, Roden says, “might have experiences so different from ours that we cannot envisage what living a posthuman life would be like”. Watts’ novel imagines, and narrates, just those conditions that are not communicable to us either subjectively or objectively.

Consider, for instance, Susan James, the linguist in the Theseus crew. She has been charged with figuring out how to communicate with the
aliens. Susan is one of those people who have done “deliberate violence to their own minds”, dividing their brain into “four fully-conscious hub personalities and a few dozen unconscious semiotic modules, all working in parallel, all exquisitely carved from the same lump of gray matter”. Susan and her alters – Sascha, Michelle, and Cruncher – exhibit the condition that was known in the 20th century as Multiple Personality Syndrome. Back then, Siri reminds us, this condition was either pathologized, or else deemed to not really exist. “People were fucking barbarians about multicores back then – called it a disorder, treated it like some kind of disease. And their idea of a cure was to keep one of the cores and murder all the others”.

In the late 21st century, in contrast, multiple personalities can be “induced deliberately” through elective surgery. This is what Susan has done, Indeed, she says, “a modern brain can run dozens of sentient cores without getting too crowded. And parallel multitasking has obvious survival advantage”. Susan even speculates that ancestral human beings might have had multiple cores, and that “our integration may have actually occurred quite recently” in evolutionary terms. And she notes that “some experts think we can still revert to multiples under the right circumstances”, even without surgery.

Another form of posthuman splintering can be found in Amanda Bates, Theseus’ military officer. Her body is cranked up with “overclocked reflexes and carboplatinum augments”. She also has large numbers of drones (battle robots) under her command; they work as physically detached extensions of her own nervous system. With Bates’ enhancements,

you can drop instantly into the sensorium of anyone under your command, experience the battlefield from any number of first-person perspectives. Your every soldier is loyal unto death, asking no questions, obeying all commands with alacrity and dedication to which mere flesh could never even aspire.

However, for the most part “Bates only ran her drones; she never inhabited them”. For they are extensions of her autonomic sensory-motor system, more than they are of her higher consciousness. Just as we walk without having to consciously consider the mechanics of walking, so the drones operate on their own, outside of consciousness, managing their own moment-to-moment movements. Like walking, all these autonomic
processes go on much faster and more efficiently when they are not subject to any sort of conscious attention. Amanda’s conscious mind, therefore, functions largely as a “bottleneck”. Her job is to keep the drones on a short leash, and prevent them from running amok. Their “electronic reflexes” are “slaved to” her much slower “meat reflexes”. This allows her to delay or cancel their self-initiated actions.

Amanda Bates’ relation to her drones mimics and amplifies the way in which, more generally, consciousness relates to more basic nonconscious processes. As Siri puts it:

Make a conscious choice. Decide to move your index finger.
Too late! The electricity’s already halfway down your arm.
Your body began to act a full half-second before your conscious self ‘chose’ to, for the self chose nothing; something else set your body in motion, sent an executive summary – almost an afterthought – to the homunculus behind your eyes. That little man, that arrogant subroutine that thinks of itself as the person, mistakes correlation for causality: it reads the summary and it sees the hand move, and it thinks that one drove the other.

This passage refers to famous experiments originally carried out by Benjamin Libet. He discovered that “readiness potentials” for a chosen action build up in our motor neurons before we are even aware of making a choice. In other words, we decide first, and have a conscious awareness of deciding only a half-second later. The “working mind” that actually makes decisions is “something else”: it is part of my brain, but it cannot be identified with the conscious “I”. When I think that I am choosing a particular course of action, I am really just ratifying a decision that has already been made on a nonconscious level. The “homunculus” of the conscious mind doesn’t actually determine anything, according to Libet; at best, it may have the power to veto previously-selected actions, by blocking their physical execution.

There is one more “freak” on Theseus: the vampire Jussi Sarasti, the ship’s commander. In describing Jussi, Watts ingeniously updates familiar pop-culture vampire motifs by giving them biological rationales. Vampires are an offshoot of early Homo sapiens, with stealthy habits and sharpened cognitive abilities. They “split off the human lineage something less than a
half-million years ago, and persisted (albeit in small numbers) into the beginning of historical times”. Vampires have superhuman talents, which originally evolved in order to help them outwit their human prey – which is to say, us. Among other things, they display “omnisavant pattern-matching and analytical skills”. When faced with an ambiguous visual image like the duck-rabbit or the Necker cube, for instance, vampires “can see it both ways at once”. Therefore, unlike us, “they can hold simultaneous multiple worldviews”. This means, among other things, that “vampires understand quantum physics, right down in the gut. It makes sense to them”. They are powerful problem-solvers because “there is no process by which they ‘work out’ these solutions: they simply see them, fully formed, laid out instantly. They don’t even have to think about it consciously”.

The other side of the vampires’ ability to see multiple perspectives at once is that they cannot be fooled by virtual reality simulations. With their advanced visual abilities, they skip right past the content of virtual images, and instead “see the pixels” themselves. According to Metzinger, ordinary human minds regard mental representations as if they were transparent: we look right through them to the things that they are supposed to represent. This is why we can be deceived so easily by optical illusions and other tricks of perspective. Vampires, however, do not suffer from this error. They see visual representations as representations, rather than mistaking them for the things to which they refer. We might say that, where we are “naive realists”, vampires are natural-born deconstructionists.

The vampires do have one fatal flaw, however. It is an ironic side-effect of their extreme visual acuity. Because of the ways that their visual neurons are cross-wired, they suffer “grand mal-like feedback seizures” whenever they see right angles. This explains (without reference to Christianity) why vampires, as folklore insists, are afraid of crosses. The crucifix glitch was not a problem for them in prehistoric times, because “you don’t find many right-angles in nature”. But once baseline human beings started building houses and other structures with right angles, the vampires couldn’t cope. Thanks to this “one stupid linked mutation”, they quickly went extinct.

Although vampires disappeared at the start of recorded human history, their genes were not lost. Rather, these genes “are widely spread among the population; they’re just dormant in most of us”. This makes it possible for biomedical corporations to bring vampires back to life in the late 21st century. The idea is to exploit their extraordinary cognitive skills. The
resuscitated, posthuman vampires are given medication – “anti-Euclidean Neurotropes” – to suppress the crucifix glitch. Supposedly their dependence upon these drugs will also prevent them from turning against us.

By ordinary human standards, of course, vampires are entirely sociopathic – even when kept on a “very short and unbreakable leash” due to their dependency upon human-supplied medication. After all, “if the only thing you can eat is your own kind, empathy is gonna be the first thing that goes”. But the corporations insist that this is not a problem, given that sociopaths are among the most successful in business, industry, and medicine; ruthless pragmatism, lack of conscience, and freedom from fuzzy touchy-feely emotions like empathy are prerequisites of success in today’s corporate environment. In fact, corporations themselves, as legal entities, meet all the diagnostic criteria for clinical sociopathy under the DSM-IV.

In short, vampire behavior is not much different from that of corporations – which are already recognized as “persons” under the law. The genetic tweaks that revive vampires, like the augmentations that produce posthuman skills, are logical steps for a society that is relentlessly obsessed, above all else, with extracting profits, maximizing utility, and optimizing cognitive processes.

I haven’t said anything yet about Siri Keeton’s own posthuman enhancements. In fact, Siri is the weirdest one of all: the most radically denatured human character in the novel. In childhood, Siri suffered from epileptic convulsions. He was cured by the removal of one whole brain hemisphere. The remaining hemisphere was able to “take up the slack”, at least as far as cognitive operations are concerned. But cognition isn’t everything. In all sorts of other ways, Siri became “a different person than the one who used to occupy this body”. In effect, the original Siri was “murdered”; and another entity “grew back out of what was left”.

The new Siri finds that his emotional life has been destroyed. He no longer feels things in the way that he used to. Nothing makes intuitive sense to him any longer. Instead, he is forced to rebuild his understanding of the world, and especially of human society, from scratch:
I was still working up the algorithms to get it back, still learning by observation… I learned to fit in. I observed, recorded, derived the algorithms and mimicked appropriate behaviors.

Siri’s approach to his own life resembles the way that video game players learn how to master games by figuring out the algorithms that drive the game engine.

Siri literalizes the common caricature of the geek who is far out along the autism spectrum, and who has creepy habits and no social skills. Since he is the novel’s only narrator, it takes us a long time to realize how strongly the other members of the crew dislike and distrust him, not to mention that his interpretations of the events he witnesses and describes to us are often disturbingly off. Siri comes across to other people, at best, as “disconnected”. Sometimes he is mentally paralyzed, unable to fulfill what he knows to be social expectations, because “I just couldn’t find an algorithm that fit”. By his own admission, none of his behavior is “heartfelt”. Even at his most successful, Siri is only going through the motions. He isn’t insincere, so much as clueless. As one of his few friends, exasperated by his behavior, tells him at one point: “it’s not so much that you don’t mean any of it. It’s more like you don’t know what any of it means”.

What Siri finds lacking in himself, above all, is any sense of empathy. This brings him – as he comes to realize – alarmingly close to the vampires. “In some sick surrealistic way”, he says, “I had more in common with Sarasti than I did with any human”. Siri is not a predator, but he operates from the same “distance – that chronic sense of being an alien among your own kind – ” as the vampires do. When a friend compliments him on having “reinvented empathy, almost from scratch”, after being robbed of it by the surgery, Siri responds that it doesn’t work for him in quite that way:

“I just observe, that’s all. I watch what people do, and then I imagine what would make them do that”.
“Sounds like empathy to me”.
“It’s not. Empathy’s not so much about imagining how the other guy feels. It’s more about imagining how you’d feel in the same place, right?”

Pag frowned. “So?”
“So what if you don’t know how you’d feel?”

Empathy means putting yourself in the place of someone else, so that you are able to feel their situation from the inside. But Siri is unable to do this. He never feels emotional situations internally – not even his own. He understands affective reactions from the outside, as purely formal processes. Without empathy, Siri is forced to run models and deduce algorithms. He tries to “imagine” people, things, and situations from the outside, on the basis of their surfaces alone. This becomes a recurring motif of the novel:

1. “Imagine you are a machine”
2. “Imagine you’re a different kind of machine”
3. “Imagine an artefact that embodies the very notion of torture”
4. “Imagine you are a prisoner of war”
5. “Imagine you are Amanda Bates”
6. “Imagine you are a synthesist”
7. “Imagine you’re a scrambler”
8. “Imagine that you encounter a signal”

And even, first at the start of the novel, then in the middle, and once again at the very end: “imagine you are Siri Keeton”.

Siri with his imaginings is in much the same situation as Thomas Nagel’s hypothetical observer, who wonders what it is like to be a bat. Nagel insists that this question cannot be answered by means of empathy. Human beings cannot put themselves in the bat’s place, because it is “a fundamentally alien form of life”. Nagel categorically rejects the strategy of finding out “what it would be like for me to behave as a bat behaves”. For the bat’s perceptual experience is so different from ours that “there is no reason to suppose that it is subjectively like anything we can experience or imagine”. The point is rather to try to determine “what it is like for a bat to be a bat”. We must neither assume that the bat is just like us, nor claim that, because it is so different from us, it has no inner life at all.

In order for us to discover what being a bat is like, Nagel suggests, we will have to resort to
a new method – an objective phenomenology not dependent upon empathy or the imagination… its goal would be to describe, at least in part, the subjective character of experiences in a form comprehensible to beings incapable of having those experiences.

The problem with Nagel’s “new method”, of course, is that he never explains how it might work. On the face of it, objective phenomenology is an oxymoron. “Objective” implies a third-person account, an entity examined from the outside; while “phenomenology” implies first-person experience, as it is lived on the inside. The difficulty is how to reconcile the two. Ian Bogost tries to answer this dilemma by revising Nagel’s project into what he calls alien phenomenology. For Bogost, our possibilities are limited: “the subjective character of experiences cannot be fully recuperated objectively, even if it remains wholly real”. Therefore, he says, “in a literal sense, the only way to perform alien phenomenology is by analogy”. Graham Harman, similarly, suggests that the only viable approach to an alien entity is one that “alludes to the object without making its inner life directly present”.

Bogost and Harman both provide aesthetic responses to Nagel’s dilemma. For Bogost’s analogy and Harman’s allusion do not claim to reconcile first-person phenomenological introspection with third-person objective observation and scientific experimentation. If anything, they suggest that such a reconciliation is impossible. Instead, Bogost and Harman offer approaches that are irreducible alike to first person identification and to third person verification. These analogies and allusions are not empirically testable; but they also cannot be determined by means of reason, intuition, or eidetic reduction. Rather, they unfold in an aesthetic dimension: one that is neither scientific nor strictly philosophical, and that is oblique to both subjectivity and objectivity. Such an aesthetic approach is programmatically that of science fiction.

Bogost’s and Harman’s approaches also have a close affinity with Siri’s method of “imagining”. Nagel, as I have quoted him above, uses “imagination” as a synonym for “empathy”. But Siri uses the word “imagine” to mean what he is forced to do, precisely because he cannot empathize. He must try to “imagine what would make [people] do” what they do, since he cannot tell what their actual internal feelings and desires would be like – not having any of his own. That is to say, Siri “imagines”
other entities precisely when – and because – he is incapable of approaching them intimately. Where empathy is first-person (“what it would be like for me to behave as a bat behaves”), and scientific observation is third-person (objective measurements of the bat’s behavior), Siri’s method is literally second-person (“imagine you are…”). He reconstructs people’s experiences algorithmically, precisely because he is himself “incapable of having those experiences”.

Analogy, allusion, and algorithmic reconstruction are all aesthetic operations – which means that they take place at a distance. Such processes are necessarily disinterested (Kant), and vicarious (Harman) – and perhaps also disconnected (Roden). Speculative extrapolation is a risky practice, as it cannot offer any assurances as to its own reliability. And yet, such extrapolation is not entirely arbitrary; it must start with a cogent rationale, and it must display a certain consistency in its development. Nonetheless, its accuracy cannot be guaranteed. Speculative extrapolation cannot entirely be contained within Kant’s limits of the understanding, or Sellars’ “logical space of reasons”. For aesthetic judgment, as Kant says, “is not a cognitive judgment (neither a theoretical nor a practical one), and hence it is neither grounded on concepts nor aimed at them”. “This means that speculative extrapolation is irreducibly fictional and aesthetic. As it is nonconceptual, it can neither be contained in propositions inferred from experience and from observable facts, nor deduced through conceptual analysis. The very process of speculation, which Kant outlawed in the First Critique, comes back surreptitiously in the Third.

Nagel says that I cannot empathize with a bat, or put myself in its place, because its inner experience is so different from my own. Roden, as we have seen, similarly suggests that posthuman entities might well be so different from us that we cannot empathize with them either. But for Siri, it is not a question of similarity or difference. He cannot empathize with other people, let alone alien life forms, because he cannot even “empathize” with himself. In other words, Siri doesn’t know what he himself feels and thinks. He needs to use his method of second-person modeling in order to figure out what it is like to be himself. Siri has no privileged access to his own inner experience; he does not know his own mind by immediate intuition.

Such a situation is obviously quite extreme. But it is less unusual than it might seem to be at first glance. To have a pain, or a flash of insight, or a burst of anger is one thing; to know that you have it is quite another.
Wittgenstein gets at this distinction with his notorious aphorism: “it is correct to say ‘I know what you are thinking’, and wrong to say ‘I know what I am thinking’”. The second person here takes precedence over the first. I can come to know objective facts, just as you can. But I do not need to know what I am thinking – or even that I am thinking – in order to think it. I think first, in any one of a large number of ways; knowing what (or even that) I think comes afterwards, if at all. Often, when a thought is deeply buried – when it is autonomic, or “repressed” in the psychoanalytic sense, or simply unattended to – I may never come to know that I am thinking it.

Gilbert Ryle similarly argues that self-knowledge is no different in kind from knowledge of other people: “the sorts of things that I can find out about myself are the same as the sorts of things that I can find out about other people, and the methods of finding them out are much the same”. Emotions and feelings tend to sneak up on me. I come to realize that I am angry, or self-centered, or in pain, by observing my own bodily states and actions – just as I come to realize that you are angry or self-centered or in pain by observing yours. This is more or less the case for everyone, since (as William James was the first to suggest) “we feel sorry because we cry, angry because we strike, afraid because we tremble”, rather than the reverse. The embodied response comes first; the consciousness of just what we are feeling only comes later, if at all. But Siri lives this condition in a more concentrated and exacerbated form than is the case for the rest of us. He never truly feels his own emotions; he can only infer them or imagine them.

In the world of *Blindsight*, Siri’s extreme self-alienation is as much a special talent as it is a disability. Indeed, it rescues him from the otherwise unavoidable baseline human condition of boredom and unemployment, and places him in the crew of *Theseus*. For Siri is a Synthesist. This means that he is able to “translate” the specialist languages of extremely complex scientific disciplines into terms that are accessible to others. Scientific research regularly “explor[es] terrain beyond the limits of merely human understanding”. Machines and augmented human intelligences come up with scientific results that are quite literally incomprehensible to baseline human beings. When the scientists and the AIs speak, we cannot understand, let alone judge, what they are saying; we just “have to take their word on faith”.


This is where people like Siri come in. Synthesists take a complex, multidimensional discourse – one that they themselves are unable to understand – and “use information theory to flatten it for you, to squash the tesseract into two dimensions and the Klein bottle into three”. In other words, Siri is just a “conduit”. He renders the science indirectly, by simplifying and reducing it. He uses the same tools that he does to make sense of other minds, or of his own: analogy, allusion, and algorithmic modeling. He is not fazed by discourses that are strictly speaking incomprehensible, because he already “find[s] pretty much everyone else incomprehensible, too”. In fact, the less that Siri understands of these complex discourses, the better he is at translating them: “you don’t have to feel motives to deduce them, it’s better if you can’t”.

Siri also tells us that his particular sort of alienation “came in especially handy when the real aliens came calling”. However weird Siri and the rest of the posthuman “boatload of freaks” on Theseus might seem to be, their strangeness pales into insignificance in comparison with the scramblers. These aliens from another star system are not only far more technologically advanced than we are, but also radically different from us. We do not know where they came from originally; when we meet them, they are stationed way out in the Oort Cloud, where their spaceship Rorschach orbits a large “dark” planetary body, intermediate between a gas giant and a brown dwarf. Siri admits that, compared to the scramblers, he and his crewmates remain fundamentally human: “the presence of this new outgroup had forced me back into the clade whether I liked it or not; the distance between myself and the world suddenly seemed forced and faintly ridiculous”. Blindsight’s vision of strange posthuman transformations is only a backdrop to, and a teaser for, its far stranger story of First Contact with “real aliens”.

The scramblers are quite alien indeed. As Watts says in his “Notes and References” to the novel, “I am weary of humanoid aliens with bumpy foreheads, and of giant CGI insectoids that may look alien but who act like rabid dogs in chitin suits”. Instead, Watts devises aliens who remain “biologically plausible”, in that they can be accounted for in evolutionary terms, but whose physiology, metabolism, and psychology are radically different from anything we know on Earth. For instance, the scramblers are large anaerobic (methane-breathing) organisms, whereas terrestrial anaerobes are unicellular. Also, the scramblers reproduce asexually, in apparent symbiosis with their not-quite-living spaceship. Rorschach
provides them with “intricate topographies of radiation and electromagnetic force”, which “mediate and regulate a good chunk of scrambler metabolism”. Their functioning actually requires radiation levels that would quickly kill any form of terrestrial life. In addition to all this, scrambler nervous systems “timeshare. Their sensory and motor plexii double as associative neurons during idle time, so every part of the system can be used for cognition when it isn’t otherwise engaged. Nothing like it ever evolved on Earth”. This is what gives the scramblers their enormous cognitive powers, far superior to ours.

The scramblers are so alien that – unlike most science fiction extraterrestrials – their differences from us cannot be mapped onto racial, gender, or sexual differences among human beings. They vaguely resemble enormous starfish, but they “become more alien the closer you look at them”. They do not fit into any of our pre-existing categories. No wonder the posthuman crewmembers of Theseus have such a hard time understanding them. It is not for want of trying. First, they open radio communication with the aliens. Then, they board Rorschach and have a look around. Finally, they manage to kidnap two scramblers, and take them back to Theseus for extended study. Yet all of these attempts fail. It takes them a long time to realize that, the more they try to penetrate the scramblers’ defenses, the more the scramblers have instead penetrated theirs.

Susan and her alters converse for hours over the radio with Rorschach. It seems like the aliens have monitored enough Earth transmissions to be able to speak English and other human languages. They even use colloquial phrases like “welcome to the neighborhood” and “anytime between friends, right?” But the conversation never becomes specific. Whenever Susan asks for information, she gets the runaround: “for four hours [Rorschach] managed to avoid giving a straight answer on any subject beyond the extreme inadvisability of closer contact”. Much of the alien discourse seems friendly; but if Susan pushes hard enough, she gets threats in response. Finally, Susan and her alters figure out that their Rorschach conversation partner “doesn’t have a clue what I’m saying”, and indeed “doesn’t even have a clue what it’s saying back’. The voice from Rorschach is actually something like a chatbot – or better, a Chinese Room.

The reference here is to John Searle’s famous thought experiment. Searle tells us that he doesn’t speak or read Chinese. He asks us to imagine
him locked inside a room. Through a slot in the door, he receives papers written in Chinese. Using a database of rules, he determines which Chinese ideograms to write in response, and passes his answers out the door. From an outside perspective, a grammatically correct and meaningful conversation in Chinese is taking place. But Searle himself has no idea what any of it means. As Siri summarizes the argument,

patterns carry their own intelligence, quite apart from the semantic content that clings to their surfaces; if you manipulate the topology correctly, that content just – comes along for the ride… you can use basic pattern-matching algorithms to participate in a conversation without having any idea what you’re saying.

The Chinese Room scenario has been widely discussed by analytic philosophers of mind. Much like the story of Mary, it is really a science fiction narrative in disguise. The room with Searle locked inside is just like a computer that operates entirely on the basis of algorithms. In this way, the Chinese Room is really a riposte, or a counterexample, to the Turing Test. In the latter, a human observer asks questions of an unseen interlocutor, who may be either another human being or a machine. If a machine is able to convince the observer that it is human, then we can conclude that the machine is at a human level of intelligence; we can rightfully say that it understands the conversation. Turing proposes this method precisely because the general question as to whether machines can think is “too meaningless to deserve discussion”. A concrete test obviates any need to consider what Turing calls the “mysteries” (or what today is more commonly called the “hard problem”) of consciousness. Behavior is all that matters; inner intentionality is irrelevant. As Turing puts it, “I do not think these mysteries [of consciousness] necessarily need to be solved before we can answer the question” of machine intelligence.

Searle, however, devises a scenario in which a conversation passes the Turing Test – and yet one of the partners demonstrably does not understand anything that is being said. The Turing Test, therefore, does not give adequate proof of real intelligence. Inside the room, all Searle has to do is locate the “squiggles” (sic; as he calls them) of written Chinese in a database, and respond to them with other “squiggles” that he doesn’t understand. Like a digital computer, he only has “pattern-matching
algorithms” – syntactical rules – without any semantic knowledge of Chinese. Searle concludes from this that syntax by itself cannot possibly generate, or account for, semantic content. Even if algorithmic computational machines – for which the Chinese Room is a stand-in – can participate in a conversation that seems meaningful from the outside, they are not, and cannot ever be, truly cognizant or intelligent.

However, other philosophers have argued as Siri himself does in the novel – that Searle’s thought experiment involves a “fallacy”. This is because even if the man in the room doesn’t understand Chinese,

the system understands. The whole Room, with all its parts.

The guy who does the scribbling is just one component. You wouldn’t expect a single neuron in your head to understand English, would you?

The intelligent entity who understands Chinese, in other words, is not Searle by himself, but the entire room, considered as a system. Yet note that posing the question in this way leads to a shift in focus. The room as a whole “understands” Chinese, but most people would not be willing to claim that it is thereby conscious. On the other hand, we presume that Searle is conscious, even though he does not understand Chinese. The Chinese Room argument equivocates over these differences. A single neuron doesn’t understand English; but as far as we know, it is not conscious either. Substituting an entire human being for a neuron, however, skews our intuitions; since Searle is conscious, we automatically assume that he ought to be the one who understands Chinese as well.

But is consciousness a necessary condition for intelligence, knowledge, and linguistic competence? Blindsight proposes that perhaps it is not. Siri first suggests this by describing himself as a Chinese Room. In his work as a Synthesist he operates entirely on the basis of syntax (or “pattern-matching algorithms”) without semantics; that is how he is able to translate, and convey to others, discourses that he himself does not comprehend in the least. Siri does his job pretty much automatically, in the same way that all of us do habitual tasks. The fact that he is also conscious doesn’t enter into it. Indeed, Siri argues that it is better this way, because he is not supposed to have a point of view at all. “Synthesists don’t have opinions on the job” he says; “it keeps observer effects to a minimum”. Siri knows that he will fail
at his task of translation if he becomes entangled with the system that he is trying to observe.

From a pragmatist or behaviorist point of view it should not matter whether the room (or the man inside the room) really understands Chinese or not. The conversation carries on, regardless. But Searle insists that we cannot just “black box” the question of consciousness; whether or not the Room is conscious matters, or makes a difference. At some point – and despite its linguistic accuracy – the Chinese Room will behave differently from a sentient interlocutor. Searle does not say just what the observable difference will be – but *Blindsight* offers some suggestions.

In a brilliant and closely argued analysis of *Blindsight*, Adam Glaz claims that the novel’s portrayal of *Rorschach* as a successful chatbot or Chinese Room is not credible. This is because *Rorschach* “lacks precisely the things that constitute the very heart of language: conceptualization, cognition, and (human-like) experience”. Given “the semantic, symbolic nature of language”, the scramblers could never learn to speak as they do on the basis of syntax or pattern-matching alone. In other words, Glaz makes the same underlying argument as Searle: syntax by itself cannot generate meaning. Or, as Susan puts it in the course of the novel, “pattern-matching doesn’t equal comprehension”. But Searle concludes from this that an algorithmic machine cannot be regarded as intelligent even if it does pass the Turing Test. Glaz concludes instead that such a machine could never actually pass the Turing Test in the first place.

In other words, Glaz’s rejects the thought experiment at the heart of *Blindsight* on the grounds that it is *a priori* impossible. But I think that Glaz’s is wrong, and that the novel does hold up as a speculative fiction. There are several reasons for this. In the first place, we simply do not know how far we will be able to go with nonintelligent software. How much linguistic complexity can be generated by syntactic and algorithmic means alone? The question of how far a nonconscious simulation of speech can go is something that will only be resolved empirically. Of course, at the present moment, artificial intelligence has not gotten very far. At present, as Alva Noë notes, “even the simplest forms of life – the amoeba, for example – exhibit an intelligence, an autonomy, an originality, that far outstrips even the most powerful computers”. And yet we have machines that, despite lacking any cognizance of what they do, are able to win chess games, drive cars, verbally answer research questions, and prove mathematical theorems.
There is lots of room for improving chatbots and digital assistants, even if Searle, Glaz, and Noë are right that they can never become truly sentient.

In the second place, Glaz’s *a priori* claim falls afoul of Roden’s disconnection thesis. Glaz’s grounding assumption is that any other intelligent species must be constrained by the same sorts of transcendental conditions of possibility as we are: “Rorschach’s command of syntax means that it either has typically human cognitive abilities or that it has evolved in ways characteristic of other organisms on earth”. But Roden argues that we cannot constrain the nature of posthuman intelligent entities on the basis of human norms: “transcendental philosophy should not be regarded as setting bounds on posthuman possibility”. In fact, the possibility space for intelligent nonhuman entities is entirely “anthropologically unbounded”. We cannot assume “a common world – a world as a shared horizon” between ourselves and hypothetical intelligent aliens. If this is the case for what Roden calls “wide human descendants”, it applies *a fortiori* to intelligent entities without any human precursors at all. Glaz is correct to argue that, since the scramblers are so unlike us, they should not be able to speak in the ways that we do. But he is wrong to infer from this that they would not be able to *simulate* human language, without understanding it, through other means. Such a possibility can only be tested empirically, rather than excluded *a priori*.

In the third place, Glaz spends so much time arguing that *Rorschach* could not possibly fake linguistic competence that he never notes that the scramblers’ linguistic imposture ultimately fails. It takes a while, but Susan and her alters eventually catch on. They note that the speech of *Rorschach* is excessively evasive: “It’s not just slippery, it’s downright *dyslexic* sometimes… And it mixes up its pronouns”. Much like chatbots today, it tries to cover up its lack of understanding by repeating vague generalities, insisting on arbitrary points, and throwing its interlocutors’ questions back to them instead of answering. *Rorschach* also breaks down in confusion when it is presented with “calculated ambiguity”, or with too many levels of recursion. In other words, and despite initial appearances, the scramblers do *not* have limitless linguistic ability. There are plenty of human linguistic subtleties that they cannot mimic or respond to appropriately.

This gets us to the heart of the novel’s argument. What goes for language also goes for consciousness in general. The scientists on *Theseus* ultimately discover not only that the scramblers’ use of language is fake,
but more broadly that – despite their frighteningly high level of intelligence – they are not conscious at all. The scramblers are simply not cognizant of themselves as selves or subjects. Susan discovers this when she runs tests on captive scramblers, in order to get a sense of how their minds work. She invents an iconic communication system that they apparently understand. But when she asks a scrambler to choose, from a list of icons, everything that is present in its room, it answers incorrectly. It fails to include itself in the list. There seems to be a difference between what the scrambler knows, and what it is “aware of”. It knows objectively that it is in the room, just like the various things it can see; but it is not aware of its own presence in the way that we would be, or in the way that it is aware of the other objects.

As Susan explains to Siri:

These scramblers – they know the answers, Siri. They’re intelligent, we know they are. But it’s almost as though they don’t know they know, unless you hurt them. As if they’ve got blindsight spread over every sense.

The qualification “unless you hurt them” refers to the way that Susan interrogates the scramblers. She zaps them with high-voltage microwaves whenever they either answer her questions incorrectly, or refuse to answer at all. As Siri sardonically notes, “this is how you communicate with a fellow intelligence: you hurt it, and keep on hurting it, until you can distinguish the speech from the screams”. Nonetheless, Susan and Siri try to convince themselves that what they are doing is acceptable, because the scramblers do not actually feel the pain: “you can’t torture the nonsentient”. But this is little more than a convenient alibi: “the oldest trick in the Torturer’s Handbook… never humanize your victims”. For the scramblers exhibit the same aversive behavior in response to pain as conscious organisms do. Indeed, if this were not the case, torturing them would not be effective. Perhaps it is best to say – however oxymoronic this might sound – that the scramblers do in fact feel pain; it’s just that they aren’t aware of feeling it.

The scramblers are effectively nemocentric (“centered on nobody”), to use a term favored by Thomas Metzinger and Ray Brassier. For Metzinger, a nemocentric system “would, while still being a functionally centered representational structure, not be accompanied by the phenomenal
experience of being someone”. Or, as Brassier puts it, paraphrasing and extending Metzinger, a nemocentric intelligence

would satisfy a sufficiently rich set of constraints for conscious experience without exemplifying phenomenal selfhood. It would quite likely remain functionally egocentric… but it would remain phenomenologically selfless… such a system’s reality-model would be richer in informational content than our own, because at every stage of processing, more information about earlier processing stages would be globally available for the system as a whole.

This indeed rightly describes the scramblers’ mode of being. With “blindsight spread over every sense”, they effectively have access to all the data of phenomenal, conscious experience, but without the experience itself getting in the way. Because their existence lacks any subjective dimension, they suffer neither from internal conflict nor from sentiment. They cannot be fooled – as we unavoidably are – into imagining that their own subjectivity is in any sense substantial. And their access to mental “processing stages” that we can never grasp helps them to effortlessly outthink the “omnisavantic” vampires, let alone baseline human beings.

All this is exemplified at one point in the novel, when the explorers from Theseus run into a group of scramblers tearing apart and engorging one of their own number. At first, the human observers surmise that there is “some kind of civil war going on” among the scramblers. But eventually they realize that this is “not civil war”, but rather a “data dump. Passing information”. No feelings of rivalry or hostility are involved. The scramblers’ knowledge and memories are encoded in their neurons. Usually, “each scrambler acts as a node in a distributed network”, passing the information along. But when the network is disrupted, they resort, as it were, to “sneakernet”: physical transfer of the information. All the scramblers’ actions are pragmatic and strategic; they are never held back by sentimental or personal concerns.

With their freedom from mammalian self-consciousness, the scramblers actually are – as we surely are not – what Brassier calls “rational agents operating in the concept-governed space of reasons”. Unlike us, they do not need to worry about “distinguish[ing] the normative realm of subjective rationality from the phenomenological domain of conscious experience” –
since evolution has not burdened them with the latter condition in the first place. They are entirely free of the illusions of folk psychology. The scramblers cannot be distracted by sentience; and so they have a level of sapience that we will only be able to equal in the unlikely event that we were able to achieve complete insight into the workings of our own minds. As Brassier puts it:

The nemocentric subject of a hypothetically completed neuroscience in which all the possible neural correlates of representational states have been identified would provide an empirically situated and biologically embodied locus for the exhaustively objective ‘view from nowhere’.

First with its accounts of strange posthuman alterations, and then even more with its vision of the scramblers and their nemocentric intelligence, Blindsight unsettles our anthropocentric assumptions about the value and centrality of our own consciousness. In comparison to the scramblers’ strategic abilities, “the best that consciousness can do, when left on its own” is quite feeble indeed. For too long, we have assumed that our consciousness makes us the crown of creation. And yet, as Siri remarks at one point, all our scientific studies and philosophical arguments about consciousness have simply missed the point. All those theories, all those drugdreams and experiments and models trying to prove what consciousness was: none to explain what it was good for.

The answer to this latter question would seem to be: not much. The novel suggests that phenomenal consciousness is an ultimate disadvantage in the struggle for survival; and that our success with it may well be an evolutionary accident. We have been lucky enough (up until now) to live in a sort of evolutionary backwater, in which we have not had to face competition from nonconscious intelligent organisms. Yet once we get beyond our own planet, we may well discover that
evolution across the universe [is] nothing but the endless proliferation of automatic, organized complexity, a vast arid Turing machine full of self-replicating machinery forever unaware of its own existence.

Most likely, we are in the position of “flightless birds lauding our own mastery over some remote island while serpents and carnivores [wash] up on our shores”.

Indeed, even here on Earth, we may well already be in process of weeding out conscious awareness. The posthuman modifications of the Theseus crew are already steps in this direction. Siri half-remembers the “vibrant” feelings that were his, before the operation that removed half of his brain. Compared to that “different, younger person” he used to be, he says that today he can “barely feel anything”. Siri also speculates that, even if vampires are “sentient to some degree”, their “semi-aware dream state [is] a rudimentary thing next to our own self-obsession”.

But it isn’t just Siri who speculates about this. The biologists also have their suspicions. Szpindel notes that, if human criteria were used, “we might call Rorschach a clinical sociopath”, in the same way that we can say this about corporations. And Cunningham reflects that

it [is] interesting to note how many sociopaths show up in the world’s upper echelons… How ruthlessness and bottom-line self-interest are so lauded up in the stratosphere, while anyone showing those traits at ground level gets carted off into detention.

After all, corporate executives are most successful when their competitive drives are unhindered by empathy, remorse, or self-consciousness. Since “natural selection doesn’t care about motives”, a good enough mimicry will likely be successful. Corporate executives will themselves gradually take on the characteristics of the companies they serve. As Cunningham puts it:

At least one thing an automaton lacks is empathy; if you can’t feel, you can’t really relate to something that does, even if you act as though you do.
When asked straight out if he is really saying that “the world’s corporate elite are *nonsentient*”, Cunningham hedges a bit: “maybe they’re just starting down that road”. But the direction of their development is clear.

This is why Siri is worried, at the very end of the novel, that there may be nobody left to take an interest in his news about the scramblers. If the only broadcasts he hears involve “traffic control and telemetry”, and if music has been replaced by the “familiar clicks and pops” of vampires on the prowl, this implies that the extermination of consciousness in the human population is already well under way. *Rorschach* has been destroyed, and “if the scramblers follow the rules that a few generations of game theorists have laid out for them, they won’t be back”. But it turns out that their threat wasn’t even necessary in the first place. Human beings are doing a good job of eliminating consciousness all by themselves. With a nonsentient Earth population, Siri says, “there won’t be any basis for conflict”.

It’s quite ironic, really. Siri’s experiences on *Theseus* are harrowing and traumatic. As a result, he loses all his pretensions to objectivity as a Synthesist, and learns that “point of view *matters*” after all. He is finally forced to *feel*, as well as know. He comes to recognize that

people aren’t *rational*. You aren’t rational. We’re not thinking machines, we’re – we’re feeling machines that happen to think.

All in all, Siri becomes “Human again”, for the very first time since his radical surgery “murdered” his initial self – and just as the very category of the “Human” is in process of being wiped out, replaced by nemocentric intelligence. This is why Siri at the end of the story may well be, as he fears and suspects, an anachronism: “the only sentient being in the universe”.

The real question, of course, is whether the difference between sentience and nonsentience even matters – and if it does, how? As long as the Chinese Room performs its translation task accurately, is it of any concern whether or not there is an actual consciousness inside? If consciousness is a “bottleneck”, then wouldn’t we do much better if our intelligence were “unhampered by self-awareness”? If all the activities of sentient people can be simulated undetectably by “a smart automaton… unaware even of its own existence”, then doesn’t this make phenomenal experience itself entirely superfluous? Whatever it is that we do with *qualia*, couldn’t we do it even better if they didn’t get in the way?
Cognitivist philosophers argue that consciousness must be adaptive, because it would not have evolved otherwise. They make many suggestions – none of which are particularly convincing or plausible – as to what its use or purpose might be. But perhaps they are mistaken to think that consciousness has a useful function in the first place – let alone to claim that “consciousness cannot be separated from function” (to cite the title of an article by Michael Cohen and Daniel Dennett).

But *Blindsight* suggests, in contrast to this, that consciousness is *dysfunctional*, a “creaking neurological bureaucracy” that is always getting in the way. It does not help us very much in the struggle for survival. Indeed, consciousness imposes heavy costs upon any organism that has it. In times of danger, “advanced self-awareness is an unaffordable indulgence”. And even at the best of times, “I wastes energy and processing power, self-obsesses to the point of psychosis”.

Consciousness therefore is not a necessity, but rather a superfluous luxury, an expression of “expenditure without return” (Bataille), or of “biological exuberance” (Bagemihl). It is a matter, not of utility, but of *aesthetics*, as Siri suggests:

Feedback loops evolve to promote stable heartbeats and then stumble upon the temptation of rhythm and music. The rush evoked by fractal imagery, the algorithms used for habitat selection, metastasize into art. Thrills that once had to be *earned* in increments of fitness can now be had from pointless introspection. Aesthetics rise unbidden from a trillion dopamine receptors, and the system moves beyond modeling the organism. It begins to model the very *process* of modeling. It consumes ever-more computational resources, bogs itself down with endless recursion and irrelevant simulations. Like the parasitic DNA that accretes in every natural genome, it persists and proliferates and produces nothing but itself. Metaprocesses bloom like cancer, and awaken, and call themselves *I*.

Consciousness is a bit like those extravagant features of certain animals that Darwin attributes to *sexual selection*. The peacock’s splendid tail reduces its chances of survival, but increases its probability of attracting mates, and therefore having offspring, before it gets prematurely killed by a predator. Sexual selection is often, as in the case of the peacock, driven by runaway
feedback loops that push it to ever-greater extremes. Consciousness is a similarly self-amplifying process, even though it is not linked to sexual difference. Aesthetic pleasures that are not “earned in increments of fitness” can nonetheless flourish whenever one is not under immediate threat. As Kant says, a hungry person will eat anything; it is “only when the need [for food] is satisfied” that people will distinguish flavors in an expression of culinary taste.

Siri’s account of aesthetics thus resonates with Kant’s; but it also involves an inversion of Kant’s argument. For Kant, it is only because we have conscious, rational minds that we are able to enjoy aesthetic experiences. But according to Siri’s speculation, it is irrational aesthetic experience that comes first, and that eventually generates consciousness as its byproduct. As Watts puts it in his “Notes and References” to the novel:

Aesthetics seem to require some level of self-awareness in fact, the evolution of aesthetics might even be what got the whole sentience ball rolling in the first place.

Aesthetics is very much a matter of sentience, rather than sapience. On the other hand, algorithmic, computational systems do not need an aesthetic sensibility. Indeed, they function all the better without it: just as Siri did before he was rehumanized. Big Blue could never have defeated Kasparov if it had been weighted down, as he was, with recursive self-consciousness and an aesthetic appreciation of beautiful chess moves. It is not surprising that Brassier, who idealizes nemocentric rationality, and separates agency from phenomenal selfhood, says also that he is “very wary of ‘aesthetics’: the term is contaminated by notions of ‘experience’ that I find deeply problematic”.

Nemocentric intelligences – like the scramblers, or like presumptive posthuman beings – can only find aesthetic and affective expression to be meaningless gibberish at best, and an aggressive provocation at worst. This is what Siri figures out when he imagines what it is like to be a scrambler devoid of what-is-it-likeness:

Imagine you’re a scrambler.
Imagine you have intellect but no insight, agendas but no awareness.
Your circuitry hums with strategies for survival and persistence,
flexible, intelligent, even technological – but no other circuitry monitors it. You can think of anything, yet are conscious of nothing. Imagine that you encounter a signal. It is structured, and dense with information. It meets all the criteria of an intelligent transmission. Evolution and experience offer a variety of paths to follow, branch-points in the flowcharts that handle such input. Sometimes these signals come from conspecifics who have useful information to share, whose lives you’ll defend according to the rules of kin selection. Sometimes they come from competitors or predators or other inimical entities that must be avoided or destroyed; in those cases, the information may prove of significant tactical value. Some signals may even arise from entities which, while not kin, can still serve as allies or symbionts in mutually beneficial pursuits. You can derive appropriate responses for any of these eventualities, and many others.

You decode the signals, and stumble:

*I had a great time. I really enjoyed him. Even if he cost twice as much as any other hooker in the dome–*

*To fully appreciate Kesey’s Quartet–*

*They hate us for our freedom–*

*Pay attention, now–*

*Understand.*

There are no meaningful translations for these terms. They are needlessly recursive. They contain no usable intelligence, yet they are structured intelligently; there is no chance they could have arisen by chance.

The only explanation is that something has coded nonsense in a way that poses as a useful message; only after wasting time and effort does the deception becomes apparent. The signal functions to consume the resources of a recipient for zero payoff and reduced fitness. The signal is a virus.

Viruses do not arise from kin, symbionts, or other allies.

The signal is an attack.

And it’s coming from right about there.

Siri thus comes to realize that, without knowing it, “we attacked the scramblers before *Theseus* launched. Before Firefall, even”. Kant says that
aesthetic beauty requires “purposiveness without an end”, or “purposiveness without a concept”. But in present-day terms, a message or signal without a purpose, one that is “structured intelligently” and purposively, but that is not at all useful, so that it “consume[s] the resources of a recipient for zero payoff and reduced fitness” such – a message can only be what we call spam. It is ultimately impossible to distinguish between spam and aesthetic expression.

Ken MacLeod makes a very similar point in his space opera trilogy *Engines of Light*. In these novels, the dominant lifeforms of the Galaxy are superintelligent asteroids, each of which is, in effect, a silicon computer of immense processing power. These beings are described as being like Lucretian gods, calmly pursuing their own interests, and most of the time not in the least concerned with what human beings and other sentient species do. The “gods” are endlessly diverse, but they do have “a pretty much unanimous view on one thing. They don’t like spam”:

As far as they’re concerned, we are great lumbering spambots, corrupted servers, liable at any moment or any megayear to start turning out millions of pointless, slightly varied replicas of ourselves. Most of what we’re likely to want to do if we expanded seriously into space is spam. Space industries – spam. Moravec uploads – spam on a plate. Von Neumann machines – spam and chips. Space settlements – spam, spam, spam, eggs and spam.

In other words, for MacLeod’s “gods”, as for Watts’ scramblers, our very *species-being* (as Marx calls it) is a provocation. Aesthetics is spam; spam is like a virus; and we are that virus. Our thoughts, our bodies, and our very lives are “needlessly recursive” and wasteful. Sensibility, awareness, and aesthetic enjoyment are costly luxuries in a Darwinian “war universe”. Consciousness, aesthetics, and unadaptiveness or dysfunctionality go hand in hand; and this, rather than any supposed achievements of sapience, would seem to be what distinguishes and defines Earthly life, including ours.
Thinking Like a Slime Mold

What is it like to be a plasmodial slime mold? How does such an entity think and feel, and encounter the world? The organism known as Physarum polycephalum is so strange that it seems like an alien life form from a science fiction novel. And yet it lives and grows all over the world. Plasmodial slime molds – also known as “true” or “acellular” slime molds, class Myxomycetes or Myxogastria – split the difference between unicellular and multicellular forms of life. In the primary portion of its life cycle, a plasmodial slime mold has millions of nuclei; but they are all contained within the membrane of a single enormous cell. As the organism grows, mitosis occurs, as it does in all eukaryotes: the nucleus, with its chromosomes, splits and replicates itself. But in slime molds – unlike either unicellular or multicellular organisms – this doubling of the nucleus is not accompanied by cell division. The single cell body just expands; the nuclei all divide, rather uncannily, in sync with each other.

In this stage, Physarum polycephalum is a colorful mass of protoplasm, big enough to be visible to the naked eye. It is something like a tiny version of the science fictional monster The Blob. It thrives in dark, warm, and moist locations. It moves – at the rate of something like a centimeter per hour – by sending out tendrils and flowing itself along. In this way, it explores its environment, searching for food. It eats bacteria, fungi, and decaying organic matter. When it discovers food sources, it simply engulfs them within its cytoplasm. Rhythmic pulsations continually traverse the slime mold: this is how chemical signals, as well as nutrients, are communicated from one part of the cell to another.

When it is ready to enter into its reproductive phase, the slime mold moves to a more exposed and better-lit location. It then grows numerous fruiting bodies, which are sort of like globules at the end of tiny stalks. The fruiting bodies send out spores into the world at large. The spores take on either of two forms, depending on their immediate circumstances. In drier places, they become amoebas, which move around by oozing from place to
place. But if there is enough water they become swarm cells or flagellates: each cell has a little flagellum or “tail” at one end, that it swishes in order to propel itself through the water. There are no genetic differences between these two types; as the environment changes, an individual cell can easily switch back and forth between them.

These spores are the product of meiosis, or cell division without replication. They are haploid, with only one copy – rather than the usual two – of each chromosome. That is to say, they are gametes, or sex cells. But they are not divided into two sexes – sperm and egg – in the manner of animals and plants. Rather, they belong to one or another of several mating types – a system that is also used by many species of fungi. There are at least four mating types in Physarum polycephalum. Mating types – unlike sperm and egg cells – are not specialized. They all pretty much look and act the same. Their only function seems to be to promote outbreeding. A cell of each mating type can conjugate with a cell of any of the other types, but not with another cell of its own type. When two haploid cells of different mating types fuse together, they reconstitute the full organism: a zygote, with a diploid nucleus. Once this happens, the slime mold assumes its plasmodial form. It feeds and grows. Nuclear division (mitosis) once again takes place without cell separation. The cycle continues.

One of the oddest things about Physarum polycephalum is that, in its plasmodial form, it cannot quite be defined either as a single individual (like most unicellular organisms), nor as a superorganism composed of multiple individuals (in the manner of coral reefs, anthills, and beehives). It also lacks the differentiation into tissues and organs that characterize most multicellular organisms. Instead, the slime mold is an oddly decentered entity: a collective without individuals, without any specialized parts, and without any sort of articulated (or hierarchical) structure. As the researchers Tanya Latty and Madeleine Beekman put it, even though “the behavior of the [slime mold as a] collective is a result of the behavior of its underlying parts”, these parts do not have any separate existence of their own:

Each slime mould is made up of many tiny pieces of slime mould, each oscillating at a frequency determined partly by the local environment, and partly by interactions with adjacent oscillators such that each oscillator can entrain those close to it.
Indeed, if you try to divide up a slime mold, you will discover that the “plasmodial fragments become fully functional individuals” in their own right, “within minutes of being separated from the main cell”. In other words, whenever you separate out a part of the collective, all you get is another collective. Physarum polycephalum is intrinsically — as the word polycephalum (many-headed) indicates — an agglomeration of many nuclei. In consequence, “owing to the slimy nature of acellular slime moulds”, they can only be understood as populations. Perhaps we can regard them as Deleuzian multiplicities, which

must not designate a combination of the many and the one, but rather an organisation belonging to the many as such, which has no need whatsoever of unity in order to form a system.

Physarum polycephalum has become a model organism for biological research — albeit on a smaller scale than fruit flies, mice, C. elegans nematode worms, or E. coli bacteria. In the second half of the 20th century, research mostly focused on working out the details of the slime mold’s life cycle and mode of reproduction. But more recent studies, since the year 2000 or so, have focused instead on the organism’s behavior in the plasmodial stage — and especially on its feats of learning and computation. Physarum does not have a brain or a nervous system; and yet it displays remarkable cognitive abilities. It is very good, for instance, at solving mazes. Put a slime mold at one end of a maze, and a tasty food source (usually oat flakes) at the other end; the slime mold will explore the maze, and discover the shortest path to the food.

How is this possible? The slime mold works by trial and error. It sends tendrils through all the passages of the maze. But then it withdraws from any dead ends. At the end of the process, its filaments trace out only the most direct route through the maze. In order to remember where it has gone, the slime mold “uses an externalized spatial memory system by depositing behind it a trail of extracellular slime”. These trails indicate passages that are devoid of food; the slime mold knows not to waste time and energy exploring them again. Physarums spatial memory works, not by internal representation, but rather by a physical marking of the very space that is being remembered. In this case, the map actually does coincide with the territory.
Physarum thus offers a simpler instance of what, in the case of human beings, has been called the *extended mind*. According to extended mind theory, cognition does not take place only in the brain, but involves the “coupling of biological organism and external resources”. My mental acts extend well beyond the limits of my own brain and body. I may remember your phone number in my mind; but I may also “remember” it by calling you on my iPhone – which has the number listed in its own “memory”. In effect, my phone does the remembering for me. Despite the scare quotes that I have used, there is ultimately not much difference between these two alternatives; in both cases, a piece of information is physically stored in such a way that I can easily retrieve it at will. My cognitive abilities are enhanced when I am able to offload some of the computational work onto external devices, or onto the environment.

Bernard Stiegler has long argued, along these lines, that technology is best understood as a prosthetic replacement (or Derridean *supplement*) for a uniquely human “default of being”. We are born weak and helpless, with a primordial deficit, which we overcome with the help of our inventions and machines. These include everything from the taming of fire, to the making of clothing and weapons, to the latest information technologies. Our common deficit, and our common endeavors to overcome it, are the basis of human sociality, according to Stiegler.

But Muriel Combes suggests, in her critique of Stiegler, that he wrongly neglects “the possibility that *humans share more than default* or lack”. Scarcity is an imposed condition; our basic social experience is one of an overflowing surplus (as Marx and Bataille both argue, each in their own way). Prosthetic technologies are therefore best understood, not negatively, as compensation for something that is missing, but rather positively and affirmatively, as basic life expressions, and as exuberant forms of invention, expansion, and transformation.

In any case, such prosthetic extensions are by no means limited to human beings – as even *Physarum polycephalum* shows. Couplings with, and exploitations of, external resources are central to all biology. An organism can only survive, flourish, and expand its reach by means of circuits that extend outside its body and through its environment. In this sense, the activity of thinking – including memory and cognition – is not all that different from other life activities, such as eating, breathing, and reproducing. As John Searle puts it:
Consciousness is a biological phenomenon. We should think of consciousness as part of our ordinary biological history, along with digestion, growth, mitosis and meiosis.

None of these life processes are purely internal and self-contained. All of them require a certain degree of reflexivity. And all of them require the organism to reach out to, and to collaborate with, the physical environment.

*Physarum*’s prosthetic use of external memory traces is (presumably) much less complicated than my prosthetic use of the memory of my computer; but it is still sophisticated in its own right. It involves, among other things, processes of recognition and a certain ability to make distinctions. Under normal circumstances, the slime mold avoids, not only the trails that it has itself marked, but trails marked by conspecifics as well. At the same time, it pays no attention to slime trails marked by other species. However, “if the organism perceives the presence of food” – probably by sensing chemical gradients – then “it will enter this area even if extracellular slime is present”.

All of this suggests, as the researchers tell us, that “extracellular slime does not serve as a repellent but as a cue that an area is most likely devoid of food sources”. In other words, the slime trail works as a source of information, not as a repulsing barrier. *Physarum* “uses a hierarchy of rules”; it differentiates one case from another, rather than just responding to fixed signals with equally fixed reflexes. It even shows nuance in the way that it evaluates the information it gathers.

In addition to its external spatial memory, *Physarum polycephalum* also has an internal time sense: “an intracellular memory that allows the organism to anticipate periodic events”. In order to test this ability, researchers first placed *Physarum* in its preferred warm and moist conditions. Then, at regular intervals, they exposed it to cooler and drier air for a brief time, returning it to the more comfortable conditions after each trial. *Physarum* responds to cold and dry conditions by slowing down its activity; it speeds up again when moisture and warmth are restored. After as few as three trials, the slime mold learned to anticipate the periodic return of cooler, drier air; it spontaneously slowed down once the right amount of time had passed. This is a conditioned reflex, much like that of Pavlov’s dogs; once again, *Physarum* exhibits it without the need for neurons, or for any cellular differentiation whatsoever.
The researchers who performed this experiment hypothesize that, in the absence of a brain, “the versatile rhythmic sense of Physarum stems from many different biochemical oscillators… operating at a continuous range of frequencies”. In other words, the polyrhythmic pulsations continually running through Physarum’s cytoplasm work as a kind of internal clock. “Rhythm” here is not just a fanciful metaphor. One group of researchers actually succeeded in converting the slime mold’s “electrical activity corresponding to different physiological states” into music. They were even able to manipulate the organism’s behavior in such a way as to produce whatever particular sounds they desired.

This experiment was not a pragmatic one, but just a proof of concept. The slime mold itself was presumably unable to hear, and not at all interested in, the music that was extracted from its behavior. But the very possibility of cross-species signal transduction suggests that there is a commonality – or even what Kant calls an aesthetic “universal communicability” – of rhythmic pulsation. All life seems to participate in what Steve Goodman calls “an ontology of vibrational force”, characterized by what Whitehead calls continuing “throbs of emotional energy”. In this regard, it is probably no accident that Physarum feels the passage of time internally, even though it can only recall spatial configurations externally. This precisely what Kant (and Bergson as well) would have expected, had they known that slime molds were sentient in the first place.

With all these cognitive capacities, slime molds are not only able to solve labyrinths or mazes, but also to work out more complicated computational problems, including network design and the balancing of different sorts of nutrients. Physarum can actually optimize network patterns, “maximising transport capacity of the network and minimising the size and length of the network”, in a way that is highly robust (resistant to failures and breakdowns). In one famous experiment, Physarum polycephalum was induced to reproduce the design of the Tokyo metro. Researchers placed food at points corresponding to all of the trains’ stops on a map; the slime mold worked out the most efficient pathways between the multiple food sources:

Overall, we conclude that the Physarum networks showed characteristics similar to those of the rail network in terms of cost, transport efficiency, and fault tolerance. However, the Physarum
networks self-organized without centralized control or explicit global information by a process of selective reinforcement of preferred routes and simultaneous removal of redundant connections.

Subsequent studies have repeated this success by getting slime molds to simulate other complex structures of routes and connections, including many actually existing transit networks around the world, and even the ancient Roman network of roads in the Balkans. The crucial point is that *Physarum polycephalum* accomplishes this without central direction, and without advance planning. As the researchers note, human builders generally use a “*top-down* design paradigm” for planning network structures, at least in the beginning even though, of course, as conditions change and the network grows, “modifications are made in an ad-hoc manner as needed”. But *Physarum polycephalum* offers, in contrast, a general, distributed, and entirely bottom-up “paradigm for self-assembly of robust networks”.

Less research has been done on *Physarum*’s response to “complex nutritional challenges”, but it seems that here, too, the organism is able to make careful and complicated choices. The slime mold grows best “on diets comprising two times more protein than carbohydrate”. When offered a variety of food sources, it selects whatever combination among them comes closest to achieving the ideal 2:1 ratio. It also responds to the dilution of food sources by spreading over a wider foraging area, making up in quantity for what is lacking in terms of quality. The researchers cannot entirely account for this behavior. However, they offer some suggestions:

How nutritional cues are integrated within the cellular matrix of a slime mold plasmodium is not known, but they seem likely to be fully distributed, involving local nutrient sensing mechanisms, movement, and growth responses.

Building on *Physarum* ’s abilities at network optimization and nutritional intake through distributed, bottom-up processes, Andrew Adamatsky and other researchers in “unconventional computing” have experimented with using the slime mold as a more general computational device. Because of its organizational efficiency,
slime mould is capable of simultaneous processing of inputs (light, chemical gradients, temperature, etc.), concurrent decision-making and distributed actuation. The behaviour patterns which ensue – which are essentially genetically encoded (innate) foraging behaviours – may be interpreted as forms of computation.

There is still no general consensus on precisely what *computation* means, and how it is accomplished by living organisms – let alone on the question of how, and to what extent, such biological activity can be compared with the way that humanbuilt electronic computers work. But although this is a crucial theoretical issue, it does not necessarily lead to any pragmatic difficulties. *Physarum*’s “genetically encoded (innate) foraging behaviours” can be used to solve problems, even though we do not entirely understand how and why these behaviors work. It is only a small step from interpreting the slime mold’s behaviors as forms of computation to actually implementing “massively-parallel amorphous computers” on the basis of these behaviors. In this way, it becomes possible to generalize beyond the slime mold’s own particular problems and concerns.

*Physarum*-based computation can be implemented in several ways. One approach is to work out algorithms that successfully model observed slime mold behavior, and then program computers with these algorithms. An example of this is Tero and his colleagues’ “Physarum solver”, a mathematical model that – like the organism itself – “is able to discover the shortest path between many points (stations) in a real-world network system”. This algorithm is derived, not from abstract mathematical reasoning, but from observing the actual “physiological mechanism of tube formation” in *Physarum*:

Tubes thicken in a given direction when shuttle streaming of the protoplasm persists in that direction for a certain time. This implies positive feedback between flux and tube thickness, as the conductance of the sol is greater in a thicker channel.

To the extent that an algorithm accurately models *Physarum*’s actual behavior – in terms of flows and thicknesses and the feedback between them – it will come up with the same computational results as those arrived at by the organism. But it is important to recognize that the “foragint
behavior” in fact comes first. The slime mold doesn’t instantiate a pre-given mathematical structure; rather, the structure itself is something that we abstract, *ex post facto*, from the organism’s concrete habits and activities.

Another approach to *Physarum*-based computation involves actually connecting a living slime mold substrate to “electrical inputs/outputs”, thereby instantiating “optical coupling in which the slime mould plays dual roles of computing device and electrical conductor”. Adamasky and his colleagues have even constructed plasmodial equivalents of “NOT and NAND logical gates” capable of performing “obstacle-free path planning” and other computationally difficult tasks. This is more than just a proof of concept. For slime mold-based computing may well be able to take into account factors or detect patterns unable to be discovered or revealed with conventional computer generated models where human biased conceptions… are a priori evident… *Physarum* machines can actually model living phenomena in a variety of ways undetected by conventional computational methods. The heuristic dynamic of the method lies in the fact that *Physarum* does not simply compute. It follows unbiased physical, chemical and biological laws which can be translated into computational language.

In other words, “conventional” computing is human, all too human. Its efficacy is limited by its inevitable anthropocentric bias. We tend to take computing procedures for granted; we assume that they are objective and deterministic, since each time that the same algorithm is run on the same data, the same results come out. But we ought to remember the old hacker saying: “garbage in, garbage out”. No computation is better, or more accurate, than its initial inputs. And indeed, it is not just inputs that are limited and inexact; the same is true for the algorithmic procedures themselves. Every algorithm encodes a particular set of assumptions. There are no grounds to assume that our own algorithms – or, for that matter, our truth procedures, or our moves in “the game of giving and asking for reasons” – are somehow fully objective, let alone comprehensive. The value of *Physarum* -based computing is precisely that this organism does not share our human biases and assumptions. It is not necessarily any more “objective” than we are; but it has *different* biases and assumptions than we
do. And thereby it might well suggest distinctions, and even algorithmic procedures, that we could never think of on our own.

It is also important to note that, as the researchers put it, “Physarum does not simply compute”. It does not follow rigid a priori rules, but rather operates according to a “heuristic dynamic”, relying upon “unbiased physical, chemical and biological laws”. What this means is that Physarum, like other organisms, takes opportunistic advantage of affordances and causal relations that already exist in its environment. Organized behaviors do not have to be elaborately encoded in the genes, or guided by explicit computations, when simple physical processes lead on their own to adaptive outcomes. For instance, bees are not genetically programmed to make hexagonal honeycombs; and they haven’t solved the algorithm that reveals that regular hexagons are the most efficient packing structure. The bees simply pack their wax structures next to one another, and surface tension produces the hexagonal shape automatically.

The same is true for Physarum polycephalum. It accomplishes its cognitive feats by taking advantage of shortcuts provided by already-existing material affordances (including the physical properties of slime trails and pulsations). This is yet another reason why we should regard thinking, not as something special, but as a particular physiological life process alongside all the others. Mental activities, no less than other sorts of life activities, are always embodied, and always grounded in a specific medium. We can, of course, abstract from such particularities, as we do whenever we simulate living processes in software. But we should always remain aware of the limits of such extrapolations, and be careful about describing such activities in the abstract, idealized terms of computation and cognition.

This is perhaps why, after listing Physarum polycephalum’s cognitive and computational achievements, Beekman and Latty add that the organism “even behaves irrationally”. Cognitive science usually assumes rationality – understood to mean efficiency, optimization, and maximization of utility – as its default, or bottom line. Of course, this is not because agents (whether human or not) are in any sense themselves knowingly rational. But the process of natural selection (and perhaps, mimicking this, the “invisible hand” of the so-called “free market”) is supposed to lead, automatically and even tautologically, to the elimination of inferior results. Efficient algorithms outperform, and outreplicate, those that are less so. In
consequence, only the most rational acts and procedures survive – regardless of individual intention and awareness.

Of course, this view is an extreme idealization. It could only work if each situation faced by an agent were entirely independent of all the rest, if the past had no influence upon the present, and if all agents had “complete information” about their environment. Since all these conditions are ridiculously counterfactual, the cognitivist assumption of rationality is itself (like most theoretical assumptions) little more than a bizarre fabulation, a fiction of sentience.

Nevertheless, full-on “irrational decision-making” is a relatively rare phenomenon in the world. It is something of which only the most complex and sophisticated systems (usually living ones) are capable. This is the startling point behind Latty and Beekman’s “rather surprising finding” that, just like organisms with brains, *Physarum polycephalum* makes faulty “economic” decisions when offered imperfect alternatives among its food sources. That is to say, slime molds – just like human shoppers in our consumer society – do not make “economically rational choices” in the way that neoclassical economics supposes, but instead use quick and dirty “comparative valuation rules” in order to decide what to do. Such rules do not maximize utility or desirability in a rigorous way, but sloppily compare the various alternatives available at a given moment.

Economic rationality, strictly defined, entails that “a decision maker’s preference for a particular option should not change when a new option of lesser value is added to the choice set”. But the relatively recent discipline of behavioral economics points out that human beings do not usually act this way. When comparative value rules are used, we can be misled by the addition of an undesirable “decoy”: an “asymmetrically dominated alternative” that is unequivocally worse than one of the previous choices, but ambiguous in comparison to the other. When the decoy is added to the list of possibilities, consumers tend to overvalue the alternative that is unambiguously better than the decoy. In consequence, our decision-making strategy “departs from axiomatic rationality”.

Latty and Beekman discovered that *Physarum polycephalum* is misled by the decoy effect in the same way that we are. They offered slime molds a series of infernal alternatives among “food disks that differed in the concentration of nutrients (oatmeal) and exposure to light”. The richer, more nourishing food sources were brightly illuminated (which is
dangerous to slime molds), while the diluted, less nourishing food sources were accessible safely in the darkness. Both well-fed and starving slime molds were placed equidistant from all the offered sources. Given these choices, “plasmodia made trade-offs between light exposure and food quality”. But these trade-off decisions were skewed “irrationally” when decoys were present.

Why does this happen? The researchers suggest a number of possibilities. Quick and dirty heuristics may well be more advantageous overall than rational rules, even though they sometimes lead to inferior outcomes. This is because rationality is “computationally expensive”. The organism would have to expend more energy in order to make a fully optimized choice than it would save as a result of having made that choice. Therefore,

natural selection could favour computationally efficient comparative strategies over the more accurate, but more intensive absolute decision-making strategies… Irrational behaviour can, under certain environmental conditions, be consistent with maximizing an organism’s expected pay-off.

Moreover, “initial conditions” have a powerful impact on the organism’s “final choice”; once the slime mold has committed resources to one approach, it might not be worth its while to switch to another one, even if the latter would ultimately be better. Then again, Latty and Beekman also suggest that the conditions of their experiment might have been overly artificial. The organisms’ behavior

might appear maladaptive in the context of the experiment, but may work well in the environments slime moulds have evolved in.

Most intriguingly of all, Latty and Beekman propose that “comparative decision-making strategies may arise as an unavoidable consequence of the way in which living systems process information”. In other words, biological information processing may well be subject to “intrinsic constraints” that are not comprehended by current theories of computation. The logic of a given algorithm can supposedly (or ideally) be instantiated in any medium; but in point of fact, the material properties of one or another
medium (in this case, the medium of biological cells) may limit which procedures are effective, or even possible. The quick and dirty shortcut offered by comparative variation rule may well be necessitated by underlying biological features – even if we do not know what these are.

It is also important to note that, even if Physarum’s “irrational” behavior is susceptible to mathematical analysis, the organism still acts unpredictably in any particular instance. When the same trials are run a number of times, the response is not always the same:

Even within a treatment group, slime moulds varied in their choices. This is particularly surprising as we controlled for weight, nutritional state and genetic differences.

In other words, even Physarum’s compliance with comparative valuation rules is a statistical average, rather than something observed in every instance. The authors suggest that

some of the variability we observed arises from slight differences in the experiments’ initial conditions… These small differences in initial condition, combined with feedback via biomass recruitment mechanisms, could ultimately result in the observed variability.

But isn’t this really the case with any spontaneous decision, by any organism? As Björn Brembs argues in his discussion of “free will as a biological trait”, there is a point beyond which “determinism versus indeterminism is a false dichotomy”. The “unstable nonlinearity” and sensitivity to initial conditions of biological decision-making rule out both “complete (or quantum) randomness and pure, Laplacian determinism”. As sensitivity to initial conditions approaches a point of indiscernibility, it becomes meaningless to make any distinction between “chance and necessity”. Rather, “such phenomena incorporate multiple components that are both lawful and indeterminate”. Irrational variations in behavior have a definite survival value, since they prevent predators from being able to predict and anticipate what an organism does. Indeed,
evolution has shaped our brains to implement ‘stochasticity’ in a controlled way, injecting variability ‘at will’. Without such an implementation, we would not exist.

Brembs is speaking of animal brains and nervous systems. But Latty and Beekman show that his observations apply to brainless organisms as well. For all living things, there is a margin, or a remainder, of indeterminacy in its behavior. This allows for something like a spontaneous decision. *Physarum polycephalum*’s behavior is not predetermined, but flexible and situational. As Latty and Beekman sum it up:

Slime moulds are capable of making complex decisions that integrate information from a variety of sources… This is remarkable given that the mechanism of decision-making in slime moulds must be substantially different from the neuronbased decision-making systems of animals. Our results therefore suggest that a brain is not a prerequisite for many forms of decision-making.

We can rightly say that *Physarum polycephalum* actively feels and thinks – just as it actively searches for, and consumes, food – even in the absence of tissues and organs specifically devoted to these tasks. *Physarum* does not need to have a brain, any more than it needs to have a mouth. The entire organism, in its environment, “acts as a brain” already.

Beekman and Latty suggest that *Physarum polycephalum* should therefore be more widely adopted “as a model for biological decision-making”. Not only are slime molds easy for researchers to work with; they also display their cognitive mechanisms more transparently than is the case with more complicated organisms. Biologists have mostly focused on decision-making in mammals on the one hand, and in insects – both solitary (fruit flies) and social (bees and ants) – on the other. Part of the reason for the latter choice is that insect societies “provide a nice analogy to the vertebrate brain (individual insects being equivalent to neurons, their colonies are the complete brain)”. However, the problem with any such approach remains that

we are still using individuals that are rather sophisticated neurologically and which share the same cognitive architecture –
brains and neurons as—human and other animals.

In contrast, slime molds accomplish the same processes without the “sophistication” that comes from having specialized nerve cells. They “have a radically different mechanism by which information is processed”, and thereby offer “an alternative path to complex decision making that does not rely on neuronal information processing. It seems that

very different biological systems use the same underlying decision-making processes, irrespective of the actual decisionmaking apparatus, be they neurons in a brain or oscillators in a slime mould.

Neurons and oscillators probably both work on the basis of positive feedback loops, with effects being triggered as a result of crossing various thresholds of intensity. It is true that the neural architectures of animals allow for an elaboration and amplification of these processes, in ways that other organisms cannot attain. But this is a difference of quantity, not of quality. It should not be taken to imply that non-neuronal organisms are not capable of having experiences and making decisions. The sorts of mental processes that we find in *Physarum* seem to be intrinsic to all cellular life. Slime molds allow us to observe the mechanisms of thought in something like their primordial form.

Without a neural system, *Physarum*’s thought process is evidently slow and dispersed. This may be why we find it hard to recognize. We tend to take for granted that any consciousness (or awareness, or sentience) must be unified. Kant, in particular, considers the unity of consciousness (or what he calls *transcendental apperception*) to be a necessary, transcendental condition of experience. But recent scientific research rather suggests that such unity is an empirical matter, and not a structural necessity.

As Peter Watts summarizes this research, unity of mind is apparently a function of “latency and bandwidth”. If signals don’t travel quickly enough through the brain (or equivalent), then unity cannot be maintained. This is why people whose corpus callosum has been severed have separate awareness and experience in each brain hemisphere. Further down the line, Watts suggests that the octopus may well have a non-unified consciousness, because so many of its neurons are located all along its eight arms and thousands of individual suckers. For its part, *Physarum polycephalum*, with
its millions of nuclei and no brain, is far too dispersed to have unity of experience in the way that we do, and that many animals do. The slime mold’s internal pulsations are low-bandwidth and high-latency: far slower than the electrochemical impulses that traverse the synapses of animal brains. It takes a considerable amount of time for energy, information, or nutrition to get from one part of the plasmodium to another. This limits the degree of behavioral sophistication available to Physarum. But it also allows the organism to do several distinct things at once, as its tendrils move in different directions and probe multiple food sources simultaneously.

So what is it like, then, to be a plasmodial slime mold? Physarum polycephalum offers us something like the degree zero of sentience and of decision-making. Its mode of thinking doesn’t involve concepts, or representations, or intentional objects, or self-awareness, or even an underlying unity of experience; it leaves out most of the things that philosophers have traditionally considered to be necessary or intrinsic to thought. And yet Physarum feels, and ponders, and decides. It acts in ways that are not always stereotypical, but at least to some degree spontaneous.

Physarum polycephalum continually prods, pokes, and provokes its environment. It navigates and searches, oozing and flowing and extending itself through its surroundings. And it receives responses to all these probes and movements: responses that take the form of sensations and oscillations. Physarum can do this because of the ways that it relies upon – takes for granted or takes upon faith – the affordances that its environment accords it. And it is guided by its own feelings of “adversion and aversion”. These feelings are themselves generated from its encounters with objects, fields, and energy flows all around it. And the feelings in turn generate further complex behaviors, some of which we may characterize as computation or calculation. Nonetheless, these behaviors do not require anything like knowledge of – or an intentional stance towards – the environment or any particular parts of it.

There is certainly sentience here, which involves (but which is not reducible to) survival-enhancing cognitive activities. Physarum polycephalum’s sentience is probably not a form of consciousness, at least in any manner that we could recognize. But this does not mean that it has no phenomenology, or no what-is-it-likeness. Rather, its phenomenology is dark, in the sense defined by David Roden:
Let’s call a feature of experience ‘dark’ if it confers no explicit or implicit understanding of its nature on the experiencer…

Dark phenomenology is thus intuition-transcendent…

Our access to the dark side would thus be as theoretically and technically mediated as our access to the humanly unobservable universe. The criteria for evaluating theories of dark phenomenology would presumably be those applying in other areas of empirical enquiry (instrumental efficacy, simplicity, explanatory unity within wider science).

*Physarum polycephalum* is rich in experience, although it does not know or understand this experience. For our part, we can only grasp the slime mold’s experience partially and indirectly, by its actual behavior, and by the traces of evidence (like slime trails) that it leaves behind. But ultimately this is also the case, even for the question of our insight into our own experience. We cannot “know ourselves” without appealing to forms of technological mediation. But this also means, as I have been insisting throughout this book, that sentience (or awareness of any sort) is inherently a matter of fictions and fabulations.
AFTERWORD

Twenty-Two Theses on Nature

1
We can no longer think of Nature as one side of a binary opposition. In an age of anthropogenic global warming and genetically modified organisms, not to mention Big Data and world-encompassing computing and communications networks, it makes no sense to oppose nature to culture, or a “state of nature” to human society, or the natural to the artificial. Human beings and their productions are not separate from Nature; they are just as much, or as little, “natural” as everything else.

2
We must think of Nature without any residual anthropocentrism: that is to say, without exempting ourselves from it, and also without remaking it in our own image. Human beings are part of Nature, but Nature is not human, and is not centered upon human beings or upon anything human.

3
Above all, we must avoid thinking that Nature is simply “given”, and therefore always the same – as opposed to a social realm that would be historical and constructed. Rather, we must recognize that Nature itself is always in movement, in process, and under construction. We need to revive the great 19th century discipline of natural history, practiced by Darwin, Wallace, and many others. Evolution (phylogeny) and development (ontogeny) are both historical processes; they cannot be reduced to the study of genomes as synchronic structures.

4
Nature is all-encompassing, but it is not a Whole. It is radically open. However far we go in space, we will never find an edge or a boundary. There is no way of adding everything up, and coming up with Nature as a fixed sum. There is also no way of subordinating Nature to some Theory of Everything.

5
Nature is radically open in terms of time, as well as space. The future is always contingent and unpredictable. It cannot be reduced to any calculus of probabilities. As Keynes and Meillassoux have both shown us, the future is intrinsically unknowable. It exceeds any closed list of possibilities. The radical unknowability of Nature is not an epistemological constraint; it is a basic, and positive, ontological feature of Nature itself.

In the 19th century, thinkers as different as Schelling (with his *Naturphilosophie*) and Engels (with his *Dialectics of Nature*), tried to define an overall “logic” of Nature that included – but that was not reducible to – human developments and concerns. In the 20th century, such projects were abandoned. Instead, humanity was either given a special, transcendental status (phenomenology); or else reduced to its non-organic presuppositions (scientism). Today, in the 21st century, both of these alternatives are bankrupt. We need to return to a project of thinking of Nature directly even if we reject the particular, antiquated terms that thinkers like Schelling and Engels used for their own attempts.

Schelling and Engels both tried to conceive of Nature in ways that were grounded in, but not reducible to, the best natural science of their own times. Our task today is, similarly, to conceive of Nature in ways that are grounded in, but not reducible to, the best contemporary science.

Nature is neither a plenum nor a void. Rather, conditions or states of affairs within Nature may tend either towards plenitude or towards vacancy. Usually, though, neither of these tendential extremes is reached. Things generally fluctuate in an intermediate range, between fullness and emptiness.

However, we are still on safer ground if we consider that Nature comprises something rather than nothing. We know from modern physics that quantum fluctuations happen even in a vacuum. In this sense, Nature is better understood in terms of more rather than less, or surplus rather than deficiency. Nature will never be finished, never be shaped and structured once and for all; but it has also never been “without form and void”.
Nature is not formless, and not simply homogeneous. It is rather metastable, in the sense defined by Gilbert Simondon. All-encompassing Nature is traversed by potentials and powers, or by energy gradients and inherent tendencies. At any moment, these may be activated and actualized. The most minute imbalance, or the most fleeting encounter, can be enough to set things into motion. And there is generally more to the effect than there is to the cause. The consequences of these imbalances and encounters tend to be orders of magnitude larger than the incidents that set them into motion.

The result of any disruption of Nature’s metastability is what Simondon calls individuation: the emergence and structuration of an individual, together with those of its associated milieu. Examples of this process include the precipitation of a crystal out of a solution, and the emergence and growth of distinct tissues, organs, and parts from an initially undifferentiated embryo.

Nature thus comprises multiple processes of individuation. These must all be understood in two distinct ways: in terms of energetics, and in terms of informatics.

Nature involves continual flows of energy. Energy (or, more precisely mass-energy) can never be created or destroyed, but only transformed from one state to another (the First Law of Thermodynamics). And yet this also means that energy is continually being expended or dissipated, as gradients are reduced, and entropy is maximized (the Second Law of Thermodynamics). As Eric Schneider argues, complex organized systems (from hurricanes to organisms) tend to form because they can dissipate energy more efficiently, and on a vaster scale, than would otherwise be possible. Such “dissipative systems” are internally negentropic; but this is precisely what allows them to discharge so much energy into their environments, thus increasing entropy and reducing energy gradients overall.
Today, thanks to our computing technologies, we tend to think more commonly in informational terms than in energetic ones. Physicists propose that the universe is ultimately composed of information; cognitive scientists tend to see biological organisms as information processing systems. I fear that our excessive concern with informatics has gotten in the way of a proper understanding of the importance of energetics.

15

Information, unlike energy, has no “in itself”; for information only exists insofar as it is for some entity (someone or something) that parses it in some way. This might make it seem as if information were inessential. But nothing is altogether devoid of information; for nothing exists altogether on its own, outside of all-encompassing Nature, entirely self-subsistent and without ever being affected by anything else. The transmission and parsing of information, no less than the transfer and dissipation of energy, is an essential process of Nature.

16

We might link information to perception, on the one hand, and to action on the other. Perception is how we obtain bits of information; and the parsing or processing of information issues forth in the possibility of action. A living organism gathers information by perceiving its environment; and it uses this information in order to respond flexibly and appropriately to whatever conditions it encounters. This is not just the case for animals, or entities with brains. A tree discerns water in the soil, which it draws in with its roots; it discovers insects feeding on its leaves, and releases a noxious chemical to repel them. Information processing thus mediates between perception and action.

17

Information processing involves – and indeed requires – at least a minimal degree of sentience. But we should not confuse sentience with consciousness; for the former is a far broader category than the latter. Organisms like trees, bacteria, and slime molds are probably not conscious; but they are demonstrably sentient, as they process information and respond to it in ways that are not stereotypically determined in advance. Even when it comes to ourselves, most of the information processing in our brains goes on unconsciously, and without any possibility of ever becoming conscious.
Most likely, consciousness is only sparsely present in Nature. But sentience is far more widely distributed.

18 Perception is only a particular sort of causality. When I perceive something, this means that the thing in question has affected me in some way, whether through light, sound, touch, or some other medium. But if I am affected by something, then that something has had an effect upon me. It has altered me (however minimally) in some manner or other. And this process cannot be confined just to perception. I am often affected by things without overtly perceiving them. I feel the symptoms of a cold, but I do not sense the virus that actually causes me to fall ill. I feel an impulse to buy something, because my mind has been subliminally primed in some way. I lose my balance and fall from a height, pulled by the Earth’s gravitational field even before becoming aware of it. I turn over in my sleep, responding to some change in the ambient temperature. In all these cases, something has caused a change in me; it has given rise to an effect. Information has been processed in some manner, by my body if not my mind.

19 Nature involves a continual web of causes producing effects, which in turn become the causes of further effects, ad infinitum. This need not imply linearity or monocausality: there are many causes for every effect, and many effects arising from every cause; and potential causes may interfere with and block one another. But just as energy is continually being transformed, so information is continually being processed even on what we might consider a purely physical level. This is why information, no less than energy, is a basic category of Nature.

20 Within all-encompassing Nature, the difference between the “physical” and the “mental” is only a matter of degree, and not of kind. A thermostat is, to a modest extent, an information processor; and therefore we should agree that it is at least minimally sentient – if not, as David Chalmers suggests, actually conscious. That is to say, the thermostat feels – although it does not know anything, and it is not capable of self-reflection. We can make a similar claim for a stone that falls off a cliff, or even for one that lies motionless on the ground. Gravity pulls the stone to the Earth, and the information associated with this process is what the stone feels.
Nature is not itself a particular thing or a particular process; although it is the never-completed sum, as well as the framework, of all the multitudinous things and processes – transformations of energy and accumulations of information – that take place within it. How, finally, can we characterize it? All-encompassing Nature stands apart from every particular instance. And yet it is not anything like a Kantian transcendental condition of possibility for all these instances, since it stands on the same level, within the same immanent plane, as them. Nature is neither outside history, nor the totality of history, nor a particular datum of natural or social history. It is rather what all these particular instances, all these transformations and accumulations, have in common; it is what places them all in a common world.

I will conclude by taking a hint from Alfred North Whitehead, who articulates this commonness more rigorously than I can. Whitehead translates the ancient Greek \textit{physis} not just as Nature (as is customary), but also as Process. And he equates this \textit{physis} with the narrower technical term (from Plato’s \textit{Timaeus}) \textit{hypodoche}, the Receptacle. Nature, or the Receptacle, Whitehead says, “imposes a common relationship on all that happens, but does not impose what that relationship shall be… [It] may be conceived as the necessary community within which the course of history is set, in abstraction from all the particular historical facts”.
Introduction

What is it like to be a bat?: Nagel 1991, 165-180.
Sapience (understanding, intelligence) vs. sentience (awareness, irritability, arousal): Brandom 1994, 5ff.
Wrongly attributed to Francis Bacon: Pesic 1999.
Brembs: Brembs 2015. See also Brembs 2009.
Fruit fly behavior: Maye et al. 2007.
Bacteria: Ben-Jacob et al 2006.
Slime molds: Latty and Beekman, 2010.
Eric Schwitzgebel on science fiction: Schwitzgebel 2014a.
Rehearsal: Peckham 1967 passim; and Peckham 1979, 204.
Feed forward: Hansen 2015, 141 and passim.
Spinoza on affection: Spinoza 2002, 240. Ethics, Book I, Definition 5: “By mode I mean the affections of substance, that is, that which is in something else and is conceived through something else”. Cf. Deleuze 1988, 48ff.
Kant on thoughts without content: Kant 1998, 193 (B75/A51).
Sellars, the myth of the given: Empiricism and the Philosophy of Mind, in Sellars 2012. See also Jelaca 2014.
Kant, aesthetic ideas: Kant 2000, 192.
Marder, non-conscious intentionality: Marder 2013, 153ff.
Plant sentience: Chamovitz 2013.

Chapter 1

The story of Mary: Frank Jackson’s original 1986 paper, and numerous subsequent replies to it and reconsiderations to it (including by Jackson himself), are collected in Ludlow et al. 2004.
Inverted spectra: See for instance, “The Inverted Spectrum”, Shoemaker 2003, 327-357; although the story has been discussed widely, and apparently dates back to John Locke.
Worlds in which water is not H\textsubscript{2}O: Putnam 1973.
Jackson’s change of position: Ludlow et al 2004, 409-442.
David Lewis: Ludlow et al. 2004, 77-104.
What is it like to be a bat?: Nagel 1991, 165-180.
Whitehead, we see with our eyes etc: Whitehead 1978, 170.
“There seems to be phenomenology”: Dennett 1991, 366.
Scott Bakker: Bakker 2013b.
Galen Strawson: Strawson 2010, 52.
Deleuze and Guattari: Deleuze and Guattari 1983, 18.
Kant on aesthetic experience: Kant 2000, 90 (204) and 95 (209).
Kant on intuitions and concepts: Kant 1998, 193-194 (A51/B75).
Kant on aesthetic ideas: Kant 2000, 192 (314) and 218 (342).
Kant, pure categories of the understanding: Kant 1998, 204ff (A67/B92ff).
Sellars, the myth of the given: Empiricism and the Philosophy of Mind, in Sellars 2012. See also Jelaca 2014.
As Dennett claims: Dennett 1991, 398.
Sellars, awareness of redness vs. sensations of red: Empiricism and the Philosophy of Mind, in Sellars 2012.
No “direct account of immediate experience”: Empiricism and the Philosophy of Mind, in Sellars 2012.
Anthropological science fiction: Empiricism and the Philosophy of Mind, in Sellars 2012.
Solaris: Lem 2011.
Ubik: Dick 2012.
Sellars on determinables and determinates: Empiricism and the Philosophy of Mind, in Sellars 2012.
Metzinger on phenomenal content: Metzinger 2004, 72-73, 79.
Dark with excessive bright: Milton, Paradise Lost, III 380.

Chapter 2
Broken hammer: Heidegger 1962, section 15.
Fish and water: McLuhan 2001, 175.
Bennett on the power of things: Bennett 2010, 1.
Hyperobjects: Morton 2011.
Units: Bogost 2012, 25.
Purposiveness without an end: Kant 2000, 111; 226 in original edition.
What is it like to be a bat?: Nagel 1991, 165-180.
The universe is nothing but information: Lloyd 2007; Vidral 2010.
Autopoiesis: Bryant 2011, 140 and 147.
Autopoiesis in biology: Marturana and Varela 1980.
Operational closure: Luhmann 2012.
Drive to reduce energy gradients: Schneider and Sagan 2006.
Erin Manning on autism: Manning 2013, 224-230.
Flat ontology: Delanda 2005, 58.
Consciousness cannot be separated from function: Cohen and Dennett 2011, 363.
Indeterminable and unfit for cognition: Kant 2000, 213.

Chapter 3
The Lifecycle of Software Objects: Chiang 2010.
Computers playing Go: Levinovitz 2014.
David Lewis: Ludlow et al. 2004, 77-104.
As I argue elsewhere: Shaviro 2016.
Historical route: Whitehead 1978, 64.
Science fiction and extro-science fiction: Meillassoux 2015, 5 and 36.
Conformation: Whitehead 1927, 41.
Computational modules: Pinker 2003, 40.
Bakker on heuristics: Bakker 2013a.
Use the world as its own model: Brooks 1987, 139.
Levinasian sense of obligation: Levinas 1969.
Elon Musk: Gibbs 2014.
Nick Bostrom: Bostrom 2014.
Sapience vs. sentience: Brandom 1994, 5.
Kurzweil on intelligence: Kurzweil 2006, 364.
Transcend our biology, but not our humanity: Kurzweil 2006, 136.
Self-enjoyment: Whitehead 1938, 150.
Affect cannot be independent of cognition: Zajonc 1984.

Chapter 4
Ray Brassier on the death of meaning: Brassier 2011b.
Benjamin Libet’s famous experiments: cf Libet 1999.
Daniel Dennett: Cohen and Dennet 2011; Dennett 1988.
Folk psychology: Stich and Warfield 2003, 235-255.
“Not I, but the wind that blows through me”: D. H. Lawrence, “Song of a Man Who Has Come Through”, in Lawrence 1994, 195.
Bakker on the “view from nowhere”: Bakker 2012.
William James on consequences: James 1975.

Chapter 5
C-fibers: Cf. the parodic account in Rorty 1979, 70-98.
Folk psychology: Churchland 1981, 67.
Second machine age: Brynjolfsson & McAfee 2014.
Precarity: Srnicek 2012.
Medial neglect: Bakker 2013c.
Sam Harris: Harris 2005.
Christopher Hitchens: Hitchens 2009.
Reinhardt’s subtractive aesthetic: ARTNews 2015.

Chapter 6
Logical space of reasons: “Empiricism and the Philosophy of Mind”, section 36, in Sellars 2012.
Brassier on narrative: Brassier 2011b.
Folk psychology: Stich and Warfield 2003, 235-255.
Cotard’s Syndrome: Metzinger 2004, 455ff.
Notes and References: These appear at the end of Watts 2006. An expanded version is available online at http://rifters.com/real/shorts/PeterWatts_Blindsight_Endnotes.pdf.
Roden, disconnection thesis: Roden 2014, 106; and more generally 105-123.
Vampires: in addition to the novel and its Notes and References, see Watts 2005.
What it is like to be a bat: Nagel 1991, 165-180.
Alien phenomenology: Bogost 2012.
Aesthetic judgment is not cognitive: Kant 2000, 95 (209). Wittgenstein, wrong to say I know what I am thinking: Wittgenstein 1953, 222.
Turing test: Turing 1950.
Other philosophers have argued: e.g. Dennett 1991, 435-439.
Adam Glaz: Glaz 2014.
Noë on machines vs amoebas: Noë 2014.
Roden on setting limits to posthuman possibility: Roden 2014, 88, 173.
Brassier on rationality: Brassier 2011a, 9.
Consciousness cannot be separated from function: Cohen and Dennett 2011.
Biological exuberance: Bagemihl 2000.
Sexual selection: for a recent scientific discussion, see Cronin 1993. For a speculative account, see Grosz 2011.
Kant on purposiveness: Kant 2000, 125.
MacLeod on spam: MacLeod 2000.
Species-being: Marx 1959.
War universe: the phrase is from Burroughs 1991.

Chapter 7
Speed of motion in Physarum: mentioned in passing in many studies, e.g. Saigusa et al. 2008; Ball 2008.
Parts and whole in Physarum: Latty and Beekman 2010.
Deleuzian multiplicities: Deleuze 1994, 182.
Model organism: see Beekman and Latty 2015, which pulls together and summarizes most of the research that I discuss in this chapter.
External spatial memory: Reid et al 2013; Beekman and Latty 2015.
Consciousness as a biological phenomenon: Searle 1993.
Intracellular memory: Beekman and Latty 2015.
Slime mold music: Miranda et al. 2011.
Kant, universal communicability: Kant 2000, 102ff. (218ff).
Ontology of vibrational force: Goodman 2010, 81ff.
Tokyo metro: Tero et al. 2010.
Other transit networks: Adamatsky et al. 2012
Ancient Roman routes in the Balkans: Evangelidis et al. 2015.
Complex nutritional challenges: Dussutour et al 2010.
Simultaneous processing of inputs: Mayne and Adamatsky 2015.
Algorithms derived from slime mold behavior: Tero et al. 2006; Beekman and Latty 2015.
Slime molds as biological computers: Adamatsky 2015; Mayne and Adamatsky 2015; Schumann and Adamastky 2015.
Game of giving and asking for reasons: the phrase is from Brandom 1997, 123; although he attributes it to Sellars.
Modeling living phenomena in a variety of ways: Evangelidis et al. 2015.
Bees and honeycombs: Karihaloo et al. 2013.
Slime mold irrationality: Beekman and Latty 2015; Latty and Beekman 2010.
Decoy effect: Huber et al. 1982.
Complex decisions: Latty and Beekman 2015.
Latency and bandwidth: Watts 2015a.
Dispersed consciousness of the octopus: Watts 2015b.
Dark phenomenology: Roden 2014, 85.

Afterword
Dissipating energy: Schneider and Sagan 2006.
Works Cited


Glaz, Adam (2014). “Rorschach, We Have a Problem! The Linguistics of First Contact in Watts’s *Blindsight* and Lem’s *His Master’s Voice*”. *Science Fiction Studies* 41, 364-391.


Libet, Benjamin (1999). “Do we have free will?” *Journal of Consciousness Studies* 6, No. 8–9.


Repeater Books

is dedicated to the creation of a new reality. The landscape of twenty-first-century arts and letters is faded and inert, riven by fashionable cynicism, egotistical self-reference and a nostalgia for the recent past. Repeater intends to add its voice to those movements that wish to enter history and assert control over its currents, gathering together scattered and isolated voices with those who have already called for an escape from Capitalist Realism. Our desire is to publish in every sphere and genre, combining vigorous dissent and a pragmatic willingness to succeed where messianic abstraction and quiescent co-option have stalled: abstention is not an option: we are alive and we don’t agree.