Will a Stroke of Neuroscience ever Eradicate Evil?

Ronald de Sousa

A man often believes he is leading when he is actually being led; while his mind seeks one goal, his heart unknowingly drags him towards another.

(La Rochefoucauld, Maxim #43)

Abstract: Cognitive science is widely regarded as having supplanted psychoanalysis as our most promising route to understanding the human mind. Nevertheless, both share a fundamental insight, in that each highlights, in a different way, the extraordinary extent of our self-ignorance. We may find, when we come to understand the neural mechanisms that govern our individual and collective motivations, that many of our traditional notions of "free-will" and of "good and evil" rest on systematic illusions. The neurology underlying the plight of psychopaths and of addicts will illustrate this possibility. The two cases exhibit an interesting contrast: in one case, the machinery for a certain sort of emotional learning is simply lacking, and moral emotions remain inaccessible. In the other, that same machinery has been hijacked and monopolized by the addictive substance, so that is has become effectively unavailable for further change. I argue that 'Evil' is an essentially theological notion, which dissolves once we truly jettison religious superstition, and which cannot properly be applied either to the addict or to the psychopath. While such a perspective may seem alarming at first, I shall suggest how it might provide foundations for the reconstruction of a more robust humanism, allowing for the genuine emergence of intersubjective responsibility.
For well over two millennia, philosophy has been notorious for its proclamation of metaphysical certainties, each of which contradicts all the others. In this it has almost kept up with the intricate imbecilities of religion, although philosophy can, at least, claim to have spilled rather less blood. I count myself among those that believe it is high time for philosophy to rein in metaphysical excess, to re-examine our premises and renew, as much as possible, the philosophical project in light of science. Questions about the relevance of evolutionary theory and neurosciences to ethical matters—guilt, responsibility, the nature of evil—are but part of that larger conversation about the pertinence of scientific discovery to philosophical conceptions of the world and of human nature. Conceptions of evil, liberty, and guilt are particularly well-positioned for profitable review and renewal.

Nevertheless, such a project of renewal involves inherent dangers. Chief among those is the ease with which one can be persuaded that this or that neuroscientific discovery is the one to lay waste to our existing philosophical frameworks, when in fact it merely offers a convenient illustration or a minor amendment to arguments long ago laid out by some of the more insightful philosophers in our tradition.

A striking example of this point concerns an issue that sits beyond the purview of ethics. It is one that is occasioned by the highly polemicized debate over “intelligent design”, a debate that is currently intense in the USA. The main debate is, of course, between fundamentalists and scientists. But the one that interests me is the subsidiary one that takes place among the latter, that is, among evolutionists and atheist scientists who are split over their degree of tolerance for religion. Some argue that the discoveries of evolutionary science and neurology afford clear and conclusive refutations of religious beliefs. Others argue that if we wish to promote the liberal values bequeathed to us by the Enlightenment, we oughtn't to say so even if that's true. Instead
we should take care to spare the faith of believers, by stressing that there is no strictly logical contradiction between science and religion.¹

This last assertion capitalizes on an equivocation. It is true that there may be no logical contradiction between the truths of science and the claims of religion. But it is no less true that many of the claims of different faiths are mutually incompatible, and that they are not rationally sustainable. There is no strictly logical contradiction between the truths of science and the existence of fairies, or of Santa Claus. But that gives no licence to think that it is rational to believe in Santa Claus. The relevant evidence here is empirical, not logical. Even if two positions are not literally contradictory, it is irrational to maintain both simultaneously if we have excellent reasons to consider them incompatible. The word ‘proof’ is ambiguous. In science, it refers to adequate evidence. Although the least contestable empirical proof will never be as logically watertight as a deductive one, we know that outside of logic and pure mathematics the notion of deductive proof seldom gets a grip. In all practical affairs, in every consideration of how to act, the only effective tools are empirical ones, including the elaborate consilience of growing human knowledge as a whole (Wilson 1998).

In what follows I will examine the ideas of ‘free will’ and ‘evil’, both of which, I will argue, are either gratuitous or incoherent as they are understood in common usage. Given that free will is implicated in our notions of responsibility, guilt, moral worth and demerit, the two notions are intimately linked. And although the revamping of these ideas is urgent, we will see that the hope that knowledge of what’s under the skull will suffice to effect the required conceptual reforms is only partly justified. Many of the tools needed to reconstruct these ideas (especially the idea of free will) do not come from science but from philosophical analysis, and

not particularly recent at that. (The idea of evil, conversely, was as confusing for most philosophers in our tradition as it is today in the light of common sense.) The needed reconstruction, however slight, will require great efforts, both in the conceptual and the practical sphere, and will doubtless meet with resistance in both.

Two framing ideas

To place all this in context, let me begin by sketching two framing ideas. Neither is new, but both remained speculative until recently. Now, however, they seem to me to rest on scientific foundations that render them nearly unimpeachable. These ideas are the moral irrelevance of nature, and the opacity of consciousness. The first comes from Darwin, or, more accurately, from a certain neo-Darwinism. It is the idea that nature does nothing for you and me as individuals. This may seem paradoxical at first, but the paradox will, I hope, be dispelled in a moment. As for the latter idea, it may evoke Freud, but I hope to convince you that it looks rather different in the light of contemporary cognitive science, including the neo-Darwinian idea just mentioned.

Why nature does nothing for me

To understand the meaning of the first idea, I need to explain, in a brief detour in the history of philosophy, the crucial difference between a traditional conception of function that goes back to Aristotle and Aquinas, and the modern notion of a biological function. For Aristotle, what occurs “always or for the most part” in nature provides sufficient evidence as to what nature intends. The natural teleology of any organism is part of its nature, carrying it naturally (albeit not invariably) towards the actualization of its specific potentiality. Even when an acorn is eaten before germination, it is still a potential oak. Aquinas adds a twist, which is that the actualization inherent in potentiality is mandated by God. But this difference is inconsequential. The core idea remains the same: it is the idea that certain processes are meant to
happen as a matter of natural fact, and that we can read off nature's “intentions” by carefully observing what “normally”, i.e. usually, actually happens. There is, of course, a difference between the statistical sense of ‘normal’ and the normative sense of the word, but on Aristotle’s view what is normal in one sense generally coincides with what is normal in the other. Perversions are rare, not by definition, but because God (for Aquinas) or Nature (for Aristotle) decrees is to be so. Such a decree has moral force, whatever form it takes: acting against nature is a sin, or at least an obstacle to the ideal unfolding of things. Evil is what is abnormal.

After Darwin, however, the philosophical foundation of this Aristotelian or Thomistic epistemology of values has collapsed. With the discovery of evolution, it is no longer possible to assimilate what ought to happen with what usually happens. Now that we understand the process that has given rise to the metamorphosis of our unicellular ancestors into homo sapiens, we must face the fact that what happens “always or for the most part” allows no inference as to what is of value, either to us as individuals or to our species. Aristotle's way of deciphering the book of nature makes sense only if we can assume the fixity of species. But at every stage that has taken us from unicellular organism to Homo Sapiens, one of our ancestors was a freak of nature: if all your ancestors had been normal, you'd be a bacterium.

Do we then have to conclude that teleological concepts such as that of biological function have lost their grip in the context of modern science? We do not. Analytic philosophy has devised a viable concept of objective biological function which owes nothing to statistical

2Actually, even in the Aristotelian perspective this criterion is not without its problems. Given that most acorns actually go to feed squirrels, one might be tempted to infer that feeding squirrels is their primary natural function. In the Thomist perspective, which appeals more holistically to an overall divine plan, local exceptions of this sort may be less problematic. Perhaps, then, the fact that all or most oaks produce acorns suffices to justify the overall ascription of function. But Aquinas, or rather his Vatican epigones, currently face a slightly different problem. Looking at Bonobos or Penguins, which Aquinas had no occasion to do, amply suffices to show that homosexual acts are decidedly not “against nature”. The Vatican should in all logic rehabilitate such acts, just as it has decided, with four centuries of hindsight, to absolve Galileo. No doubt, after another four centuries of due deliberation, it will.
normality, to divine planning, or even to intrinsic but impersonal “tendencies” of Nature. This is
the aetiological concept of function, which rests on the following simple and compelling idea:

To know what \( X \) is supposed to do, one just needs to know which effects, among
those that \( X \) produces, are those explain \( X \)'s presence here and now.

Like Aristotle's, the modern concept selects among the actual effects of an organ those
that it is in some sense meant to have. As in Aristotle again, this selection is based on past facts.
But unlike Aristotle we are not concerned with how often something happens. Rather, a function
is identified with those (frequent or infrequent) effects, among all those produced in the past, that
made the organ in question more ‘fit,’ or likely to get reproduced. Thus it is that the heart's
circulation of the blood, but probably not its rhythmic sound or its location left of centre, are
among its proper functions: blood circulation is why we have hearts.

This criterion is applicable in all the different cases where teleology is invoked, including
intentional actions and artifacts. If I press this or that button on my remote, it's because the effect
of doing so is what I want to achieve. If the remote is here in the first place, it's because it's been
built precisely to produce that effect. Before Darwin, it seemed plausible to say just the same of
any organ: the heart, for example, is there because it has been built by the divine engineer to
circulate the blood. We no longer need that hypothesis, but it remains true that the heart (at least
so we surmise) is there because of its contribution to the fitness of those equipped with a heart.
Specifically, it is there because organs of that sort have favoured the differential reproduction of
such organisms in the past, and thereby contributed to their presence in organism now alive.
That, then, is the rationale for identifying the circulation of the blood, and not other effects such
as the making of rhythmic sounds, as the function of the heart.
In short, we too can distinguish, no less than Aristotle, those effects that something just *happens* to have from those that it is *supposed* to have, i.e. those that constitute its *functions*. But the rationale for doing so is no longer the same.³

A crucial consequence of the difference concerns the meaning of the claim that a certain characteristic is adaptive. From that claim, we cannot deduce that the characteristic in question is necessarily beneficial to the individual possessing it. The true beneficiaries of the functions of our organs are *formal patterns of which those functions have favoured the replication*. These formal patterns are typically embodied in strings of nucleotides, but the important fact is that they are replicators, not that they are made of DNA. Other replicators are in no way excluded. This point is easily obscured by talk of the “survival of the fittest”, which appears to refer to well adapted individuals. But in truth no individual ever survives. One also tends to forget that the very term ‘sexual reproduction’ is a misnomer, in that no individual in sexually reproducing species ever actually reproduces. Bacteria reproduce, and so do sequences of DNA; thus only they, in the end, never directly individuals, can be the beneficiaries of natural selection.

On that view, should we say that:

(A) My genes are among my organs, whose functions serve my ends.

Or, on the contrary,

(B) I am an organ of my genes, serving the vestigial "teleology" of natural selection. ? Prima facie, it might go either way:

³This aetiologica conception has a long pedigree, starting perhaps with (Taylor 1964) who himself credits (Sommerhof 1950), and most finely elaborated by Ruth Millikan (1984; 1993). For debates about this conception see (Allen, Bekoff and Lauder 1998). Here is a slightly more precise analysis adapted from (Proust 1997, 216–217):

(P) An element $X$ has the "direct proper function" $F$ iff
(1) $X$ is the product of the reproduction of a previously existing element $Y$;
(2) $Y$ effected $F$ in the past owing to properties (=$i...n$) reproduced in $X$;
(3) $X$ exists because (in the causal historical sense of "because") $Y$ effected $F$
(A'). Genes are organs of the body whose function it is to produce more bodies.

Or

(B'). Bodies are organs of genes whose function it is to produce more genes.

But it can’t be the first, because, as just noted, sexually “reproduced” bodies don’t actually get replicated. It can, on the other hand, be the function of bodies to transmit genetic information, because (the information in) the gene does get duplicated. So the appropriate parallel is:

Hearts’ circulating blood is why they exist.

Bodies' passing on genes is why we exist.

More formally: The aetiological conception of function specifies:

\( F \) is the function of a type of organism or entity \( O \) iff:

* \( O \) results from the a chain of reproduction from an ancestral \( A \)

* \( A \) effected \( F \)

* \( O \) exists because \( A \) effected \( F \).

Substituting to make genes organs of bodies:

Making bodies is the function of genes, iff

* Current genes results from the a chain of reproduction from ancestral genes

* Ancestral genes effected the construction of bodies

* Current genes exist because ancestral genes effected the construction of bodies.

That is not implausible, but its plausibility rests on our being particularly vague about the identity of the bodies involved. It can be turned on its head:

Facilitating gene replication is the function of bodies, iff

* Current bodies results from the a chain of reproduction from ancestral bodies
* Ancestral bodies facilitated gene replication

* Current bodies exist because ancestral bodies facilitated genes replication.

From the point of view of causation, the two appear to be symmetrical. But if we remember that bodies are never replicated, and that only genes, as units of information, survive with sufficient self-identity through time to be recognizable, we can conclude that only replicators can be beneficiaries of selection.

Why the soul doesn't exist

In the light of these reflections it is not surprising that, contrary to what seems to be the teaching of much of modern philosophy from Descartes to phenomenology, most of our mental life is actually hidden from us. My motives, the meaning of my actions, my ‘mind’, if you will, are to a very large extent hidden from me, and will remain necessarily so forever. This thought can be summed up in the slogan that the soul does not exist. Of necessity, an immaterial and transparent soul cannot have hidden recesses.

We shall see in a moment how this opacity to ourselves applies specifically to the planning of intentional actions. But we can already see how it fits in with the previous point. If the mechanisms that govern our behaviour are not governed by a teleology that aims at benefiting the individuals as such, their true beneficiaries have little interest in revealing their true “intentions” to their unwitting agents. We are working, we might say, for the sake of aliens, who have no interest in disclosing their plans.

-----------------------------------

4Actually Descartes is far from being as unequivocal about this as he is often thought to be. In the Passions of the soul, we can find two divergent views within a page or two. In Article 26, he writes that “passions… are so close and so interior to our soul that it is impossible that we should feel them without their actually being such as they are felt to be.” Yet in Article 28 he notes that “experience teaches that those who are the most strongly moved by their passions are not those that know them best.” (Descartes 1989).
Few facts of psychology are as well established as the difficulty we find in following Socrates' counsel to “know yourself”.\(^5\) Consciousness shows us the results of mental processes, not those processes themselves. Anyone can test this for herself, just by adverting to her capacity to recall, say, the name of an acquaintance. It is obvious that the easier it is to do this, the less likely it is that we have the slightest idea how the feat was accomplished. (If one knew, one would have a method to follow on occasions when the name sought resists coming any closer than the “tip of one's tongue”.) Furthermore, numerous experiments have shown that subjects are often wrong about the reasons or motive of their own acts or mental states. In one well-known experiment, exposing men to a measure of fear caused men to find more attractive a women they had just met. But none recognized the influence of that factor. (Dutton, et al. 1974) In another experiment, Tim Wilson and colleagues asked subjects to choose an item from among a table full of what were actually identical items of apparel. Subjects would overwhelmingly confabulate reasons for preferring the ones on the right, but none acknowledged—even when debriefed and shown the evidence—the influence of position on their choice. (Wilson 2002) And since each of us has surely had the experience of catching others in a state of astounding self-deception, the inference is hard to evade that my own attempts to explain my actions may not be more reliable than those adduced by others.\(^6\)

**Two notions in need of reconstruction**

At the risk of raising hackles in some quarters, let me state as axiomatic that only suffering is intrinsically bad. Kant famously held, on the contrary, that “Nothing can possibly be conceived in the world, or even out of it, which can be called good, without qualification, except a good will.” (Kant 1959, opening paragraph) This is no place to go into that doctrine in depth, 

\(^5\)For a detailed and readable survey, see (Wilson 2002).
\(^6\)For a recent discussion of confabulation and self-deception from the neurological point of view, see (2005).
but it seems to me to run into two fatal objections, one conceptual and one empirical. The first is that it is easy to explain what is bad about a bad will in terms of suffering that it causes. If harm and suffering did not exist, nothing would distinguish a good will from a bad one. If that is right, then it is the value of what the will aims at that determines the moral worth of the will itself, and not the other way around.\footnote{Cf. (Hurka 2000). That is not to say that in our world as it is there may not be virtues that lend moral worth to their possessors, but it does seem difficult to imagine how those virtues could have acquired such value in the first place in a world deprived of any distinction between pleasure and plain.}

The second problem is the one with which I shall mostly be concerned in what follows. Under the scrutiny of science, the notion of the will seems to become complicated to the point of disintegration. But if the notion of the will has lost all clear sense, then so has Kant's maxim.

Let us then say that an agent acts in a harmful way insofar as she is attempting to cause harm or suffering. In that sense, of course, for the evildoer, doing evil has to do with motivation and intention. But in the victim, i.e. in those that undergo the harm, evil is primarily a matter of suffering. Harm and suffering in themselves pose obvious practical problems: how to avoid them, how to heal or palliate them. But they pose no intellectual or moral problem. Most suffering comes from nature, and so the issue of motivation is beside the point. Nevertheless, we seem to have an innate tendency to treat every event as proceeding from an intentional agent. We come innately equipped with what Justin Barrett has called an HADD, or “hyperactive agent detection device.” (Barrett 2004); (Dennett 2005).

This is where theology enters the picture. Evil in the full sense—Evil with a capital E—is first the gratuitous and incomprehensible cruelty of a divinity that humans seek to appease. In a Manichean dualist system, which subsists in much of allegedly monotheistic theology, it is a pure pole (sometimes embodied in the figure of Satan) that acquires its identity from its
opposition to a perfectly good deity. To explain the evolution of religion into monotheism, a psychoanalytic hypothesis retains much plausibility: why would anyone pretend to believe that the powers that govern the world are perfectly benevolent, if not precisely to appease a being that is, if not perfectly malevolent, at best an unpredictably savage and bloodthirsty tyrant? And since that tyrant reads every one of our thoughts, the only way to persuade him of the sincerity of one's abjection is first to convince oneself of it. Thus theology constructs the Problem of Evil, which stems from the fact that the empirical evidence, which tallied perfectly with the hypothesis espoused by “primitive” religions that saw the gods as wanton and cruel, appears entirely to contradict the new hypothesis of omnipotent benevolence. A new discipline, theodicy, is then charged with solving the very problem which it has just invented: namely to explain how it could be that an omniscient, omnipotent and perfectly benevolent deity could cause or allow the infliction of such calamities on its beloved creatures.

Intellectually, this is not a hard problem. Doing away with the problematic premise about the divinity will do the trick. But that does not lift the suffering, nor does it remove the power of Justin Barrett's HADD which persists in attributing whatever we undergo to the actions of some agent. In most human beings (all but mutants, perhaps?) there seems indeed to figure such an innate device, which is doubtless as useful, on the scale of evolutionary history, as the modular visual mechanisms that ground our “visual intelligence” (Hoffman 1998) but also make us susceptible to certain systematic visual illusions. In either case, knowing the facts does not always suffice to dispel the illusion.

Evil in its full sense, then, consists in the will to cause suffering, allied to limitless power. Among humans the paradigm cases of evil men are the great tyrants, Hitler, Stalin, Mao, whose power equalled their cruelty. But this grandiose conception of Evil tells us nothing about the psychological mechanisms that generate such a will to cause suffering. Such an Evil Will remains stuck on the mythical and essentially theological plane. To do Evil is to take the side of
Satan against God. Thus one of the privileged statements of what it means to choose Evil for Christians goes back to Augustine, who seems to have retained some traces of the Manichean world view to which he had subscribed before his conversion. What Augustine stresses, in telling the story of his theft of pears, is the intention to do Evil for the sake of Evil, and not for pleasure, or out of selfishness. Augustine himself found this paradoxical (Augustine 1909–14, II, x. 18).

Presumably, if the theological notion of Evil had not occurred to him, the idea of pure transgression, symbolized by the myth of Original Sin, would have made no sense at all. The intention to do Evil of which Augustine speaks seems to rest essentially on the myth of Evil itself.

Freed from theological baggage, our natural inclinations to conceptualize things in terms of Evil threaten to be more confusing than helpful. Once we give up the idea of Sin—that is, of some intrinsic Evil—as a hypothesis to explain the painful consequences of the behaviour of other people, we might hope to reconstruct a notion of evil on a less grandiose but more practical basis. We will then speak of undesirable acts, of harmful acts; of the need to prevent such acts; or to modify aggressive impulses; or protection of society, and so forth. The conception of criminal law that would correspond to these preoccupations will be concerned with prediction, prevention, dissuasion, and rehabilitation (PPDR). It will have to find some suitable accommodation with the majority opinion that those who have caused harm deserve the infliction of harm in retribution. The urge to blame may be innate, the category of Evil is a great deal broader than can be explained in terms of suffering, and its content varies considerably from one society to another. There seems to be something irresistible to common sense about the notion of sin and the institutions of interpersonal and social guilt allocation built upon it.

Does the emotion of guilt arise from a sort of illusion? Clearly it is possible to feel guilty without actually being guilty. Indeed, a certain conception of determinism is sometimes adduced in support of the claim that guilt is never legitimately ascribed. Common sense, on the contrary,
tells us that guilt is sometimes a matter of objective fact even where there no feeling of guilt. In some cases, indeed, one might be all the more guilty for failing to be aware of one's guilt. The elimination of the concept of guilt therefore remains a deeply implausible prospect. But what is the proper application of the notion of guilt? Perhaps it is indeed a purely objective notion, which should be kept apart from issues about responsibility that normally attach to it. I'll come back to this question in a moment.

Sartre or Libet?

Jean-Paul Sartre is the author of a formula almost as famous as Descartes's *Cogito*, and perhaps equally compelling: *We are condemned to be free.* Even the most ardent determinist must concede that if asked to decide to bend a finger or not to bend it, *waiting to see* is not an option. If I bend my finger, I have decided to do it. If I claim to be waiting for the outcome of the forces converging on me at that moment, I have in effect also made a decision to do nothing.

Nevertheless, Sartre's bold formula encounters some uncongenial facts of logic and experience. First, what is felt by the agent as a free act is constituted, at least in part, by the functioning of a whole set of subpersonal mechanisms, of which we are no more conscious, and of which we have no more voluntary control, than digestion or the marshalling of our immunological defences. We become aware of such processes only when one of them goes wrong, or undergoes an unaccustomed shift. Such anomalies have been brought to light by Daniel Wegner's experiments to detach the feeling of having acted from actual causal efficacy (Wegner 2002). Participant in a table-turning séance can be utterly convinced, quite falsely, that they were merely following the table's motion, not causing it to move. And conversely, using ingenious manipulations inspired by the Ouija board, Wegner was able to show that agents can
also become convinced of having both willed and produced certain effects which as a matter of fact were due to causes entirely out of their control.⁸

A classic experiment due to Benjamin Libet, the main results of which have now been convincingly replicated, is even more disconcerting (Libet 1985; Haggard 2005). What he showed was that in certain situations, such as the simple finger-wagging case above, the consciousness of deciding lags almost a whole second behind the activation of the readiness potential that signals that the machinery of motion has already been triggered, and about 200-300ms after the actual initiation of movement in the motor centres of the brain.

We are forced to conclude—on pain of giving up the axiom that causes precede their effects—that Sartre was wrong to assume that free-will could be assimilated to the causal efficacy of conscious decision. It seems that free decision, on the contrary, is no less an effect of something else than the act itself. Conscious will, in the words of Wegner's title, is an illusion.

Wegner himself adduced the hypothesis that the consciousness of having willed something is often a confabulation, that is, an explanation devised after the fact on the basis of the information that the act has actually taken place, conjoined with the conventional assumption that such and such a motive would constitute an acceptable reason for committing it. I've already mentioned the widespread occurrence of confabulation.⁹ The concept of confabulation is reminiscent of Freudian defense mechanisms, but the motivation is broader. Rather than to conceal from oneself unavowable desires, most confabulation is motivated merely by the need to explain one's own behaviour in some way or other. Some late news from neuroscience, however, appears to favour another hypothesis that is liable to restore a certain role of consciousness in the

---

⁸(For critical comments, see (Proust 2005, 50).)

⁹Confabulation comes in several different types, and there is fascinating work on the brain mechanisms behind it. See (Hirstein 2005).
elaboration of voluntary behaviour. This is the alternative hypothesis put forward by Patrick Haggard, according to which the overall planning of an act undoubtedly precedes consciousness, but concedes to the latter a monitoring and predictive role in the detail of the plan's execution: « conscious intentions are at least partly reconstructions, rather than mere reconstructions. » (Haggard 2005, 293). Nevertheless, the basic inference from the result of Libet's experiment is not impugned: an act's intention, at least in some cases, is not elaborated in the theatre of consciousness.

Contrary to Wegner's claim, however, this does not entail that free-will is altogether illusory. That inference would be warranted only if cleaves to the supposed link between freedom of choice and conscious deliberation. But there are independent reasons for doubt on that score, if only in the common experience of daily life. We perform innumerable acts without the slightest consciousness of having deliberated about them, and yet clearly without having been constrained to perform them. Libet's and Haggard's results therefore suggest that we should be looking for a reconstruction rather than the elimination of the notion of free action.

The desired reconstruction faces two tasks: first, to defuse the supposed conflict between freedom and determinism; and second, to return to Sartre's argument in the hope of finding a way to reconcile the neurological data with the subjective sense of the inevitability of choice.

The first task was already accomplished by David Hume. His analysis was the first clearly to declaw the apparent threat to “liberty” stemming from “necessity”. As Hume made clear, the opposite of liberty is not necessity but constraint. And the opposite of necessity is not freedom, but chance. Free will is not to be confused with chance, despite any attempt that might be made to confuse the issue by appealing to a certain concept of “gratuitous act” such as is illustrated in Gide's Les caves du Vatican, in which a character attempts to transcend determinism by committing a random murder.

Tourette, Alcoholics, Psychopaths
The subjective experience of freedom, whether veridical or not, does not attend every act. It appears to be notably lacking in some cases of alcoholism and drug addiction, and Gille de La Tourette syndrome (GTS). This obviously does not mean that in such cases the behaviour is determined, whereas in normal situations it is not; rather just as Hume claimed, free action requires the absence of constraint, that is, of a sort of cause external and contrary to the agent's desires. Where an impulse is both extremely intense and unrelated to the agent's own conception of her life goals, preferences and interests, it is appropriate to speak of constraint even where the causal factor is not external to the agent. One is tempted, in such cases, to speak of irresistible impulses, and that is why obscene ejaculations of GTS are viewed differently from similar behaviour coming from subjects not affected by that particular infirmity.

Actually GTS sufferers do not entirely confirm the view of their symptom that views them as involving irresistible impulses or “tics”. Phenomenologically, their impulses are felt as intense desires. But must we assume that whoever acts on a desire, in contrast with someone acting under constraint, is always capable of resisting the impulse? In a forthcoming article, Tim Schroeder concludes that the neurological analysis of desire requires us to reject any simple dichotomy between desire and constraint. The consideration in which he grounds his claims are these. Motor centres in the cortex are not responsible for the actual planning of an action. Rather these centres “propose” behavioural sequences belonging to a relatively limited repertoire at their command, and these proposed acts are then routinely inhibited by the basal ganglia. Among the functions of the basal ganglia is the integration of instructions stemming, among others, from the cognitive cortex and the motor centres. In addition, they respond to signals coming from the reward centres, which is what makes learning possible. Unless the goal of a “proposed” behaviour is endorsed by the subject, it will normally be inhibited. Occasionally, however, even in the ordinary life of a normal subject, an overlearned set of responses—such as a route long used in a daily commute—will usurp an explicit intention to go elsewhere. At the crossroad, I
should be turning right towards my new place of work; instead, I “automatically” turn left towards my old one. Yet here is no mere reflex, still less am I constrained to do this. So it is highly significant that anomalies have been noted in the basal ganglia of GTS.\textsuperscript{10} Schroeder suggests that it is plausible to infer that GTS sufferers are not acting under the influence of desires as we normally understand them. Rather, they are “moved by non psychological structures.” We are probably right, therefore, to infer that in such cases the moral responsibility of agents is not implicated in the same way as it is in cases where a choice is made, from among a array of possible alternatives, in the light of a hierarchy of plans integrated in a whole life and personality.

Drug addition has a number of characteristics in common with GTS, including the subjective feeling that the act which the agent would like to withhold is neither a simple reflex nor, strictly speaking, an \textit{involuntary} act, but that the category of the \textit{voluntary} does not quite fit either. In a 1999 survey of relevant research, Gardner et David described a number of well known experiments in which animals devoted themselves entirely to self-stimulation of reward centres in the brain, to the detriment of all else, including sex and nourishment, to the point of death from starvation. Yet the substances consumed by different addicts do not necessarily have anything in common in their stereo-chemical characteristics: “In fact the single essential commonality of the addicting drugs is the neurobiological one -- their ability to acutely enhance the pleasure/reward circuitry of the brain. ( (Gardner and David 1999, 102 ; emphasis in the original).)” And they continue a few pages later:

“at the very outset and at late stages of recovery from drug addiction, humans are perhaps less similar to laboratory animals. But during the active addictive phase,

and during both the acute and short term withdrawal and abstinence phase, we are perhaps most similar to laboratory animals.” (ibid., 127)

When the word “vice” is detached from its traditional association with moralistic disapproval, that is with the notion of Evil, I suggest that it is best defined as follows:

A vice is a mode of behaviour that one cannot forbear to indulge in, even though it no longer provides any pleasure.

As it happens, the tics of GTS and the addicts' so called “abuse” of their favourite substance fit this definition rather well. Something like vice in this sense has even been experimentally induced in rats (Berridge and Valenstein 1991). In the light of which, one continues to be astounded by the irrationality (bad faith, ignorance, stupidity, or simple cruelty) of politicians who persist in affecting to believe that it is good policy to treat addicts as criminals, in the same league as psychopaths.\footnote{An important qualification is required here. While the evidence for changes in the brain's sensitivity to chemical signals is impressive, there are also some compelling dissents. See especially (Davies 1997) (Alexander 2001; Vignemont and Singer 2006), on the importance of contextual factors in the modulation of brain processes governing empathy and addiction.}

What then of psychopaths? Work by Robert Hare (Hare and Babiak 2006), later much publicized by Antonio Damasio (1994), has brought to light a number of other anomalies in some of the persons commonly known as “psychopaths” or “sociopaths”. Prominent among these anomalies is an emotional deficit that manifests itself in various ways, notably in the absence of the somatic responses that usually attend the apprehension of suffering in others, or even of imminent pain in themselves. These subjects suffer not from irresistible inner constraint, but rather from the absence of those inner boundary constraints that in most of us restrict the range of possible behaviour. (Blair, Mitchell and Blair 2005).
Whether this deficit is best described as a lack of empathy depends on just how that is defined. Vignemont and Singer (2006) usefully distinguish between empathy and sympathy, contagion and cognitive perspective taking. True empathy is distinctive in three ways:

There is empathy if: (i) one is in an affective state; (ii) this state is isomorphic to another person’s affective state; (iii) this state is elicited by the observation or imagination of another person’s affective state; (iv) one knows that the other person is the source of one’s own affective state.

There may be research that targets the identification of neurological correlates of empathy understood in terms of those four conditions, but I am not aware of it. On the other hand, there are data that suggests that there are important consequences of that emotional deficit. This pertains to the difference between two kinds of learning, and to the lack of a capacity for negative learning in psychopaths. In this respect there is a significant difference between psychopaths and addicts, which makes it impossible to justify treating both alike. Their brains have been so altered by the drug that their behaviour cannot be modified by the usual learning process involving the reward centres of their brains: normal rewards and punishments therefore remain without effect. The case of the psychopaths in this respect is precisely the converse: internal constraint is not unbreakable but altogether absent. Thus in both cases, but for opposite reasons, the structure of their behaviour and desires is beyond the reach of punishment.

Psychopaths often seem to us to embody a clear instance of absolute Evil. And yet the sort of Evil in question here is in a sense the very opposite of the type of Evil characterized by Augustine. For Augustine, as we saw, the bad will consists in wanting to do Evil for its own sake rather than for pleasure, ambition, or any other intrinsically intelligible goal. This is not true of either the addict or the psychopath: the former is often inclined to wish she could refrain from indulging in her addition, while the different seems to be utterly indifferent to the suffering that the pursuit of present pleasure might cause others or his future self. But the two are importantly
different. The brains of addicts have been so altered by the drug that their behaviour cannot be modified by the usual learning process involving the reward centres of their brains: normal rewards and punishments therefore remain without effect. The case of the psychopaths is precisely the converse: internal constraint is not unbreakable but altogether absent. Thus in both cases, but for opposite reasons, the structure of their behaviour and desires is beyond the reach of punishment.

The word ‘punishment’ is generally taken to refer to some sort of retribution—a notion that is related to that of Evil. Abolishing that notion would not, of course, remove any of numerous other practical reasons pertaining to PPDR—protection, prevention, dissuasion, and rehabilitation—that might justify removing from society individuals whose behaviour persists in causing harm and suffering. Is there any way, beyond those practical considerations, that we can accommodate the intuition that psychopaths are “evil”?

Part of what is at stake here is the question of whether it is possible to reconcile first person responses with a third person attitude that acknowledges the causal determinism which, on the basis of the postulate that “ought implies can”, appears to invalidate the attribution of guilt. In an influential article, Peter Strawson argued that the objective point of view that sees in the behaviour of others a natural fact, subject to causes as to chance, is not incompatible with our dispositions to react to others' behaviour from the point of view of the first person.

Strawson stresses the practical and psychological impossibility of eliminating “reactive” emotions such as resentment, indignation, gratitude, vindictiveness, forgiveness, and so on. (Strawson 1962). Such attitudes signal a subject's engagement and participation in a system of personal and social relations. Strawson grants that as attitudes, these are incompatible with the “objectivising” which assumes that the behaviour of others is due to determining causes or to chance. Furthermore, the behaviours motivated by these reactive attitudes are not necessarily the ones most likely to be mandated by a preoccupation with PPDR.
Recall the Sartrean paradox. In the third person, the actions of other people can always be envisaged as stemming from chance and necessity. We can also take this stance in relation to our past selves. But it is impossible to adopt such an objective point of view on one's own present predicament: when regarding my own position, I am indeed forced to be free. We can regard Strawson's contribution as consisting in the extension of this observation about the first person to the second person. Given a certain proviso, it is not possible to regard one's interlocutor as a thing. The essential proviso is this: one's interlocutor must be a genuine participant in a relation of reciprocity. Such participation requires something like what philosophers call a “social contract”. What philosophers call a “social contract” is, of course, a fiction. There is no actual explicit contract. But what underwrites the fiction is a first person acceptance in the participants of responsibility, that is of the possibility of fault or guilt. In participating in social relations with others, I effectively claim the right to be found guilty. In this way, I take upon myself the freedom that I am “forced” to exercise.

That is, I suggest, the mechanism that allows the emergence of moral reciprocity. It leaves open the possibility that we might have to make a distinction between intentional agents that are also authentic participants in social life, and others who are precluded from such participation as a result of the disruption of certain neural circuits by chemical or environmental factors. But the rationale for making such a distinction will not be moral. Instead, it will be based on the practical requirements of PPDR.

The idea that any human being should be “precluded” from participating in the fundamental game of reciprocity is a shocking one. Douglas Heinrichs (2006) has forcefully argued that the victim of an assault or murder by a psychopath cannot reasonably be expected to regard the criminal as a mere fact of nature. The psychopath is not failing to play the fame of mutuality: he is merely playing it in a particularly hateful way. I cannot take, towards him, the attitude that would be appropriate to a tsunami. Heinrichs has further urged that we should regard
human decisions not as the outcome of simple deterministic systems, but of chaotic networks. That would make it irrational to make any predictions with a degree of certainty sufficient to justify taking any measures based solely on PPDR. We should, in other words, take account of our collective self-ignorance in a way that requires us to adopt authentically reactive attitudes to every human being.

Coda: Dismissing the bogies of Determinism and Reductionism

Let me return, before concluding, to the biological significance of our self-ignorance. Moralists and psychoanalysts stress our need to hide from ourselves our own selfish or otherwise unavowable motives. But I have suggested that most of our ignorance of our own motives has more general roots. If we keep that in mind, we might attribute our confabulations not so much to our selfishness as individuals, as to our manipulation by essentially alien factors, unconcerned with our individual interests—real or apparent.

There has been much talk in the literature of the last couple of decades about group selection, which is often credited with the otherwise paradoxical evolution of altruistic impulses in the members of a group 12. But it is important to remember that in every group of cooperating individuals free-riders and cheats have an advantage. The fitness of a gene for cooperative altruism depends on its relative frequency, specifically in comparison with cheaters. Among cheaters, cooperation can increase fitness. But in a group of altruists, cheating can succeed. In some cases the result will be mix of “evolutionarily stable strategies” (ESS) at some point of equilibrium that depends on the values of the various relevant parameters. (Maynard Smith 1984) One might speculate, as a possibility deriving plausibly from the point of view I have sketched here, that psychopaths might represent one such strategy—one that will be stable providing its

12 See (Sober and Wilson 1998), and for my own account of their book (de Sousa 1999).
relative frequency remains close to the point of equilibrium. The peculiar features of their brains that render them virtually incapable of experiencing either the emotion of compassion for others or anxiety about their own future suffering would then be functioning in a biologically normal way, differently than, but on exactly the same terms as, the brains of non-psychopaths. Nature does things perfectly—but only from the point of view of the genes responsible for psychopathic tendencies.

Let me stress, once again, that to say that psychopathy is genuinely an outcome of natural selection in no way implies that psychopathy is a good thing from the perspective of any particular individuals, group, or society affected by it. Neither does it imply that it is “good for the species”. We need to get rid of the thought that what is natural is good, and by the same token of its converse: that whatever isn't natural is bad. That inference, common as it is, is seldom more sensible than the character in an old Gardner Rea New Yorker cartoon who refuses a drink offered by an airline attendant: “No thank you,” she says, “I don't think nature intended us to drink while flying.” Nature constrains what is possible; but the fact that something is a product of evolution has no bearing on its value.

The common disparagement of “scientism” heard today may partly stem from a dim recognition that the natural world is not necessarily friendly. But two prominent forms of this sentiment are really nothing but bogeymen: “genetic determinism” and “reductionism”. To be sure, some popular formulations of the notion of gene manifest a simplistic “genocentrism”, grounded in an obsolete conception of “beanbag genetics” (De Winter 1997) Some thinkers advocate a holistic view of gene functioning based on genetic networks (Gavrilets 2004); others propose to downgrade the status of genes to a mere causal factor among others in complex cascades in immensely complex and environment-responsive “developmental systems” (DST) that constitute the true replicators. (Griffiths and Gray 1994; Oyama 2000) All these alternative hypotheses are of interest in themselves, and greatly complicate our conception of the nature and
role of genes. But none affect the philosophical lesson I am attempting to draw here. For whatever the precise details of the mechanisms that perpetuate species, selection will never favour individuals as such in sexually reproducing species. Setting aside the uniqueness of individual genomes, the importance of external factors affecting development emphasized by DST guarantees that no two organisms will ever develop into identical phenotypes. In the end, nothing is ever actually reproduced except information, whether it be embodied in DNA, in networks, in developmental systems or other carriers of ancestral patterns, but never in individual organisms as such. It follows that fitness, interpreted as the differential capacity for propagation of biological patterns, is at best only indirectly relevant to the interests of any particular individual. Thus new perspectives on genetics do not in any way undermine the framing ideas sketched above.

Furthermore, it is worth noting that most statements of biological fact relating to evolution are formulated in statistical rather than nomological terms. Any statistical statement must be interpreted differently depending on whether it applies to a homogeneous population or to a heterogeneous one. In the latter, statistical statements conceal the concrete underlying facts about individuals. To illustrate this, compare two (putative) statistical claims, one of which concerns a radioactive substance while the other purports to relate to a human population:

(i) An atom of radioactive substance R has a 68% chance of disintegrating by date X.

(2) 68% of Canadian citizens have brown eyes.

The first relates to a homogeneous population. In a mass of radioactive element, each atom has precisely the same chance of disintegrating before a given date. (If the half-life of the element is x years, then each atom has a 50% chance of disintegrating within x years.) We can say that each atom has an identical \textit{propensity}. In the case of (2), by contrast, it would be absurd to infer that each Canadian has a propensity to have brown hair. Every Canadian, barring the
intervention of hair stylists, has a propensity to have just the hair colour she has. But a given statistic concerning a heritable trait can come from any combination on a continuum that goes from 68% having propensity 1 to all members having an identical propensity 68%. The following important conclusion follows: No statistic can either confirm or disconfirm the hypothesis that there is, at the neurophysiological level, any strict determinism at work.

So much for the first bogeyman. As for reductionism, it is still capable of arousing passionate invective. Witness a review by a prominent intellectual journalist, Leon Wieseltier, of Dennett's recent book *Breaking the Spell* (Dennett 2006). Wieseltier begins by quoting Dennett:

"Like other animals... we have built-in desires to reproduce and to do pretty much whatever it takes to achieve this goal.... But we also have creeds, and the ability to transcend our genetic imperatives"

and comments: “And then more, in the same fine antideterministic vein: ‘This fact does make us different.’” Notice that the smooth passage from the idea that we are “different” to the idea of “anti-determinism” is a complete non-sequitur. In the sense intended by Dennett, “transcendence” has strictly nothing to do with determinism. Wieseltier goes on:

Then suddenly there is this: “But it is itself a biological fact, visible to natural science, and something that requires an explanation from natural science.”....

Dennett does not see that he has taken his humanism back. Why is our independence from biology a fact of biology? And if it is a fact of biology, then we are not independent of biology. If our creeds are an expression of our animality, if they require an explanation from natural science, then we have not transcended our genetic imperatives. The human difference, in Dennett's telling, is a difference in degree, not a difference in kind — a doctrine that may quite plausibly be called biological reductionism (Wieseltier 2006).
Despite the passions it arouses, reductionism is a false problem.\textsuperscript{13} Holism and reductionism are best thought of as alternative ways of focusing on the very same ubiquitous phenomena. Emergence is a fact of nature at all levels, including the relation between the properties of elements and those of their constituent atoms. The emergence of consciousness is therefore not a unique phenomenon in nature. It transcends the properties of the constituents that make it possible only in the sense that the properties of water “transcend” those of oxygen and hydrogen, or the properties of diamonds or nanotubes transcend those of the carbon of which they consist. All these can equally well be considered in the perspective of reductionism or of holism. More specifically, the elements and their new organization are causally sufficient (and sometimes necessary) for the appearance of the emergent phenomenon. Nevertheless, the emergent phenomenon is logically distinct from all the properties of the elements observed in isolation.

Biology offers many examples of levels of properties emerging holistically. Maynard Smith and Szathmáry (1995)\textsuperscript{14} describe eight “major transitions” of evolution at which a new level of organization of the constituent elements open up new possibilities. The invention of language and that of consciousness—if they don't simply form a single transition—each illustrate a transition similar in this respect, though at the opposite end of the sequence, as the origin of life or the organization of the genetic material into chromosomes. Every point of transition marks a phenomenal explosion of new possibilities.

\footnotesize
\textsuperscript{13} A drawing by Douglas Hofstadter showed the mystical word MU with letters made up of the words REDUCTIONISM, HOLISM, themselves written with letters consisting respectively of the words HOLISM, REDUCTIONISM, which in turn are spelled with letters made up of repeated small 'MU' (Hofstadter 1980, 310).

What I have been suggesting is that Strawson's conception of reactive attitudes is another such transition, which allow us to reconcile Sartre's observation with the scientific point of view that conceives of explanation only in terms of chance and necessity. We can think of this transition as marking the emergence of a fresh level of social interaction, on the basis of capacities of individuals that have been predisposed, by their biological evolution, to certain forms of behaviour and certain forms of response in the social context. Reductionism is nothing more than the essential spirit of science, which sets itself the goal of analyzing the precise condition under which a new form of organization allows a mass of known elements to generate a previously unpredictable—perhaps even previously inconceivable—phenomenon. In the present instance, what is in question is the study of the biological factors—including evolutionary pressures or drift, as well as neurophysiological processes—that might illuminate the possibility of social life itself.

One practical consequence of taking the present point of view might be a radical reform of the penal systems currently in force in so-called civilized countries. Another might be the development of programs of education that would make it easier for each individual to enter into that sphere of reciprocity which underlies individual responsibility, that “social contract” that rests on the subjective engagement of each with others.

I hope that if one adopts the perspective I have sketched one might cheerfully consent to the abolition of the ontological category of Evil. It need not be missed, so long as we preserve as intrinsic enablers of social life those biological predispositions that have induced us to reify Evil and project it into theistic mythology. A strictly naturalist perspective has nothing to fear from the notion of emergence. Conversely reductionism, which is as it were the other side of emergence, offers no threat whatever to the most inclusive and ambitions conception of human possibilities. While not infinite, we have reason to think these possibilities may be unbounded.
References


Heinrichs, Douglas. 2006. Comments on an earlier version of this paper. AAP, Washington DC.


