Will a Stroke of Neuroscience ever Eradicate Evil?

Ronald de Sousa and Douglas Heinrichs

> *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 will illustrate this possibility. We argue that 'Evil' as traditionally understood is an essentially theological notion, which dissolves once we truly jettison religious superstition, and which cannot properly be applied to the psychopath. Notions of evil, freedom, guilt, and moral responsibility must be reformulated to be compatible with our understanding of human functioning based on neuroscience and yet make sense of the complex range of our responses, individually and collectively to the psychopath as well as our intuition that he embodies absolute evil. While such a perspective may seem alarming at first, we shall suggest how it might provide foundations for the reconstruction of a more robust humanism, allowing for the genuine emergence of inter-subjective responsibility and the moral community.*

How ought we respond to the psychopath? An example from the psychiatric practice of one of us concretizes the issues involved.

Jane is an administrator with the federal government, who entered treatment at age fifty-four. She had a terrible marriage when young, divorced her husband and raised two children alone, all the while successfully advancing her career. At forty-two years of age she got romantically involved with her boss Bill, a bachelor. They eventually moved in together, and were extremely happy. She described him as the center of her life for the next 12 years. Then, at the age of 66, Bill was brutally murdered in his home during a burglary while Jane was at work. The crime was committed by Joe, a 25 year old who is unequivocally psychopathic.

As the trial unfolded Jane became more agitated and distressed. She was especially distraught by Joe's callous attitude when he talked about the killing of "the old man", his lack of feelings of remorse, and even his seeming lack of much concern about his own punishment. She had fantasies of brutally torturing him herself. He was found guilty. Jane sensed that the judge and jury were outraged at Joe's attitude and she expected the maximal sentence. But she became agitated about making a victim impact statement. She felt it was very important that she do so, yet both feared and expected that Joe would not be moved in any way. She found some consolation in believing that others present would be. She wanted them all to know how special Bill was to her.

We can all sympathize with Jane. Indeed, our most fundamental, visceral intuition is that the psychopath is the paradigmatic embodiment of evil. But as we understand more of the distinct

neurology of the psychopath, can we really hold him responsible for who he is? And if we cannot, does anyone count as blameworthy or evil? Our efforts to formulate an appropriate response to the psychopath, both as individuals and as a society, will be only as coherent as our most basic conceptions of evil, guilt and moral responsibility. Philosophers have not, alas, done a good job of clarifying those crucial concepts; they are notorious, indeed, for proclamations of mutually inconsistent metaphysical certainties. We believe it is high time for philosophy to rein in metaphysical excess, 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 demands caution. One can rush to the conclusion that this or that neuroscientific discovery is the one to lay waste to our existing philosophical frameworks, when in truth it merely offers an illustration, or a minor amendment, to long-standing philosophical insights. Sometimes philosophers did get it right.

In what follows we will contend that the ideas of ‘free will’ and ‘evil’ 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. The needed reconstruction, however slight, will require great efforts, in both the conceptual and the practical spheres, and will doubtless meet with resistance in both.

Two framing ideas

To place all this in context, let us begin by sketching two framing ideas. Neither is new, but both remained speculative until recently. Now, however, they seem to us 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 us as individuals. Any air of paradox in this assertion will be dispelled in a moment. The latter idea evokes Freud, but contemporary cognitive science, including the neo-Darwinian idea just mentioned, has placed it on more solid foundations.

Why nature does nothing for us

To understand the meaning of the first idea, we need to explain, in a brief historical detour, the crucial difference between a traditional conception of teleology and the modern understanding of 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. An acorn is a potential oak, even if it is eaten before germination. Thomas Aquinas added 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: certain processes are meant to happen as a matter of natural fact, and we can read off nature's "intentions" by carefully observing what "normally", i.e. usually, 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 the latter sense generally coincides with what is normal in the former. Perversions are rare, not by definition, but because God (for Aquinas) or Nature (for Aristotle) decrees it to be so. Such a decree has moral force insofar as acting *against*

nature is taken to be a sin, or at least an obstacle to the ideal unfolding of natural processes. *Evil is what is abnormal*.

In the light of evolution by natural selection, however, what happens “always or for the most part” allows no inference as to what is of value to us, either as individuals or as a 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 on the way 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.

Must we conclude that the teleological concept of biological function has lost its 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 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, as a result of natural selection.*

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. Instead, 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 be reproduced. Thus it is that the heart's

circulation of the blood, but probably not its rhythmic sound, are among its proper functions: blood circulation is *why we have hearts*.[1]

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.

A crucial consequence of the difference concerns the meaning of the claim that a characteristic C is adaptive. From it we cannot deduce that C 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. 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. The very term 'sexual reproduction' is a misnomer, in that no individual in sexually reproducing species ever literally *reproduces*. Bacteria reproduce, and so do sequences of DNA; thus only they, not the individuals that contain them, can be the direct beneficiaries of natural selection.

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—our motives, the meaning of our actions, our 'mind', if you will—is to a very large extent hidden from us.

---

[1]This aetiological conception has a long pedigree, starting perhaps with (Taylor 1964) who himself credits (Sommerhof 1950). It was best elaborated by Ruth Millikan (1984). For debates about this conception see (Allen, Bekoff and Lauder 1998).

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 do not intrinsically aim at benefiting individuals as such, then we are working, we might say, for the sake of aliens who have no interest in disclosing their plans.

Few facts of psychology are as well established as the difficulty we find in following Socrates' counsel to "know yourself".[2] Consciousness shows us the results of mental processes, not those processes themselves. Anyone can test this for herself, just by adverting to our capacity to recall the name of an acquaintance. 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 did know, one could apply the same method 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 motives 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 woman they had just met. But none recognized the influence of that factor. (Dutton, et al. 1974) Subjects have been shown to confabulate reasons for their choices in many other experiments. 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.[3]

Two notions in need of reconstruction

Kant famously held 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

---

[2]For a detailed and readable survey, see (Wilson 2002).

[3]For a recent discussion of confabulation and self-deception from the neurological point of view, see (Hirstein 2005).

paragraph). This is no place to go into that doctrine in depth, but it seems to us 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 its propensity to cause suffering. But if harm and suffering did not exist, nothing would distinguish the outcomes of a good will from those of a bad one. If that is right, then the moral worth of the will depends on the value of what it *aims* at, not the other way around.[4]

The second problem is the one that will mostly concern us 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.

We shall take it as axiomatic, then, that *only suffering, and not the will, is capable of being intrinsically bad*. An agent *acts in a harmful way* insofar as she is attempting to cause harm or suffering. In that sense doing evil has to do with motivation and intention. But in the victim who undergoes 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 problem. Most suffering comes from nature, and so the issue of motivation is beside the point. And we have already seen that tending to our welfare and happiness is not the function of nature.

Nevertheless, we seem to be innately equipped with what Justin Barrett has called an "hyperactive agent detection device" (*HADD*). (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, Evil is a pure pole

---

[4]Cf. (Hurka 2000). That is not to deny that, in our world as it is, virtues lend moral worth to their possessors. But those virtues could not have acquired such value in the first place in a world deprived of any distinction between pleasure and plain.

(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 manifestly is, if not *perfectly* malevolent, at best unpredictably savage and bloodthirsty? If, moreover, that deity 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 saddles itself with the *Problem of Evil*: while the empirical evidence tallied perfectly with the hypothesis espoused by "primitive" religions that saw the gods as wanton and cruel, a new discipline, *theodicy*, is charged with explaining how an omniscient, omnipotent and perfectly benevolent deity could cause or allow the infliction of 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 Barrett's HADD which persists in attributing whatever we undergo to the actions of some agent. Such an innate device is doubtless as useful, on the scale of evolutionary history, as the modular visual mechanisms that ground our "visual intelligence" (Hoffman 1998). Both mechanisms, however, also give rise to certain systematic illusions, which knowledge of the facts is not always sufficient to dispel.

But what then of cases of evil truly inflicted on us by agents, that is, by other human beings? Great evil in its full sense, derived from theology, 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. 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 his early Manicheanism. Augustine tells the story of his theft of pears, stressing his *intention* to do Evil *for the sake of Evil*, not for pleasure or out of selfishness. Augustine himself found this paradoxical (Augustine 1909–14, II, x. 18). If the theological notion of Evil had not occurred to him, the idea of pure transgression, symbolized by the myth of Original Sin, would presumably have made no sense. The evil intention of which Augustine speaks rests essentially on the myth of Evil itself.

Freed from theological baggage, our natural inclination to conceptualize things in terms of Evil threaten to be more confusing than helpful. Once we give up the idea of Sin as a hypothesis to explain the painful consequences of certain acts, we might hope to reconstruct a notion of evil on a less grandiose but more pragmatic basis. At the extreme we could attempt to eliminate the concept of evil completely, except as a shorthand term for whatever activity causes suffering. Such behaviour would be seen as presumably rooted in an individual's neurobiology, derived from genetic endowment and/or environmental influences. Bad behaviour then would be the straightforward outcome of these causal processes, not the choice of an evil will. On this broadly determinist view, guilt is never legitimately ascribed. Neither, it should be noted, is praise. This implies that HADD is a mechanism for which, in fact, no proper object exists at all, since there are no agents that choose anything. This does not preclude HADD at times being a useful fiction. At the level of social policy, this position encourages a utilitarian and pragmatic approach. 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).*

But this position feels incomplete and unsatisfying. Is there 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, and while what is regarded as

Evil varies considerably from one society to another, the scope of the concept of Evil is a great deal broader than can be explained in terms of suffering. 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.

Is it indeed possible that the emotion of guilt arises from a sort of illusion? Clearly it is possible to feel guilty without actually being guilty. Common sense, conversely, tells us that guilt is sometimes a matter of objective fact even where the feeling is absent. In some cases, indeed, one might be all the more guilty for failing to feel guilty. The elimination of the concept of guilt therefore remains a deeply implausible prospect. This reflects our deep commitment to the notion that at least some of our actions are chosen—that, in some meaningful sense of 'free', we are free agents after all. But is this intuition defensible in light of modern science? And if so, which of our actions are freely chosen? Certainly not all of them qualify. And are they coextensive with those for which we should be held responsible and found guilty if they result in suffering to others? Perhaps guilt is indeed a purely objective notion, which should be kept apart from issues about responsibility that normally attach to it. Such a view is reflected, for instance, in early Greek thought as well as much of current tort law. Are some of us freer than others? If so, how would we know? The opacity of consciousness precludes a direct answer from introspection. And yet some sense must be made of our intuition of freedom.

Sartre or Libet?

Jean-Paul Sartre quipped that *We are condemned to be free.* Even the most ardent determinist must concede that if asked to *decide* to wag a finger or not to, *waiting to see* is not an option. If I wag 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 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. These are no more conscious, and no more under 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. Daniel Wegner (2002) has shown that the feeling of having acted can be detached from actual causal efficacy. Participant in a table-turning séance can be utterly but falsely convinced that they were merely following the table's motion, not causing it to move. 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 were actually due to causes entirely outside their control.[5]

A classic experiment, due to Benjamin Libet (1985) and convincingly replicated by Haggard (2005), is even more disconcerting. What he showed was that in situations such as the 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. It also lags 200-300ms after the initiation of movement in the brain’s motor centres.

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 Wegner's word, is an illusion.

Wegner himself adduced the hypothesis that the consciousness of having willed something is often a confabulation, an explanation devised by the agent after the act, on the basis of conventional assumptions about what would constitute a plausible reason for an act of that sort. I've already mentioned the widespread occurrence of confabulation, a concept reminiscent

---

[5](For critical comments, see (Proust 2005, 50).)

of Freudian defense mechanisms, but deriving from a wider set of motivations than the latter. Rather than to conceal from oneself unavowable desires, most confabulation is motivated merely by the need to find some kind of explanation for one's own behaviour. Some late news from neuroscience, however, appears to favour another hypothesis that is liable to restore a certain role of consciousness in the elaboration of voluntary behaviour. This is the alternative hypothesis put forward by Patrick Haggard. This views the overall planning of an act as preceding 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 preconstructions, 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 elaborated outside the theatre of consciousness.

Contrary to Wegner's claim, this does not entail that free will is altogether illusory. That inference is warranted only if cleaves to the supposed link between freedom of choice and conscious deliberation. But common experience gives us independent reasons for doubt on that score. We perform innumerable acts without any consciousness of having deliberated about them, yet clearly without constraint. 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 must defuse the supposed conflict between freedom and determinism and find 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”. For the opposite of liberty, Hume argued, is not necessity but constraint; and the opposite of necessity is not freedom, but chance. Free will is not to be confused with chance. It is not incoherent then to see ourselves as free, in the sense of exercising our liberty, even though the complex organism that is exercising that freedom evolved from purely deterministic and causal mechanisms.

Freedom is *emergent*. The emergence of consciousness and of freedom are not unique phenomena in nature. The properties of water "transcend" those of oxygen and hydrogen; the properties of diamonds or nanotubes transcend those of the carbon of which they consist. In all these cases, 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 important examples of levels of properties. Maynard Smith and Szathmáry (1995)[6] describe eight "major transitions" of evolution. At each one, a new level of organization opened up a radically novel space of possibilities. The inventions of language and of consciousness illustrate a transition similar in this respect, though at the opposite end of the sequence, to the origin of life or the organization of the genetic material into chromosomes. Every point of transition marks an explosion of new possibilities.

But if freedom does not depend on the absence of necessity, what in fact characterizes it? How far can we trust our subjective experience of freedom to accurately identify acts appropriately considered free and for which we can be judged morally responsible?

Psychopaths and the neurology of freedom and constraint

The subjective experience of freedom, whether veridical or not, does not attend every act, even when the act is "free" of external constraints. Where an impulse is both extremely intense and unrelated to the agent's own conception of her life goals, it is sometimes appropriate to speak of the act as "unfree" even where the proximate causal factor is not external to the agent. This line of analysis moves the point of constraint inside the skin—or brain—of the agent. It presupposes a distinction within the person between a self and something not quite deserving of

---

[6]Maynard Smith, J., and E. Szathmáry. 1999. *The origins of life: From the birth of life to the origins of language*. Oxford; New York: Oxford University Press.

the name. On one model, the first typically involves some complex of central processes embodying abstract representations and reasoning, self awareness and assessment, integrated goals, secondary desires and deliberation; while the second refers to a series of subpersonal, non-psychological discrete modules. Freedom is lost and constraint is imposed to the extent that a module is able to override the "self" or central processing capacity. Fodor (1983) proposed a related distinction between modular and central components of mental activity. A demonstration of the application of such a model to an understanding of mental disorders and the nature of explanation in psychiatry has been recently elaborated by Murphy (2006). The precise nature of the central component, its neurological substrates and whether it lends itself to further modular characterization remain open questions.

Substance abuse is consistent with the proposed distinction between central processes and subpersonal, non-psychological modules. It creates a more difficult and ambiguous instance to test this distinction. Early on in the course of substance abuse the behaviour seems to reflect volitional choices, yet as it progresses to addiction and dependence the behaviour appears more automatic and involuntary. (Gardner and David 1999; LaLumiere and Kalivas (2008)). A similar argument has been made for the case of Gille La Tourette (GTS) (Schroeder 2005).

What then of psychopaths? Work by Robert Hare (Hare and Babiak 2006), much publicised by Antonio Damasio (1994), has brought to light a number of other anomalies in some of the persons commonly diagnosed as "psychopaths". Prominent among these anomalies is an emotional deficit that notably manifests itself 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 appear to 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. Equally important is the impact of the absence of these constraints on the moral development of the psychopath. As recently elaborated by Blair, Mitchell and Blair (2005), an

alleged deficit in the amygdala results in an impairment in the violence inhibition mechanism (VIM). Normally humans, as well as many other mammalian species, tend to refrain from activities that cause visible suffering in others and are more likely to engage in behaviours that reduce others' distress. Moral socialization is theorized to depend heavily on the pairing of the activation of this mechanism triggered by the distress cues of others with representations of the behaviour that causes the distress, i.e. moral transgressions. The psychopath lacks moral inhibitions on causing suffering to others and is willing to cause such suffering if needed to further his goals. Thus the psychopath is deterred from inflicting suffering neither by the fact that to do so induces reactive distress in him nor by the conviction that to do so is morally wrong.

At a practical level, the decision as to whether public policy should seek punishment and retribution for the offending psychopath is rendered less critical by the latter's relative immunity to punishment or deterrence. It appears that for now the best we can do is to develop policies to optimally protect ourselves. This is all compatible with the principles of PPDR. Yet at the individual and personal level we clearly react quite differently to the psychopath than to the person with GTS or to the addict. The psychopath no more chose to have a deficit in his amygdala than the person with GTS chose to have basal ganglia malfunction, and less so than the addict chose to distort the response of her reward centers by ingesting addictive substances. Yet rather than seeing him as the victim of constraints on his freedom we see the psychopath as embodying pure evil. Is this difference in our response rooted in a defensible intuition? If our working principle is that we excuse, as the result of constraint, those actions where a subpersonal module overwhelms the deliberative capacities of the person's central processing components, and hold the person responsible for actions under the control of those central processes, then the intuition about the psychopath is consistent. For his problem appears to involve not constraint, but the malformation of an important aspect of the central processing itself, i.e. moral socialization. The psychopath acts based on true deliberations but they reflect a distorted view of

others and his moral obligations to them. The psychopath in fact is the “sort of person” who inflicts suffering on others without regard for that suffering. When he does so he is “being himself” and acting “in character” in a way that the person with GTS and the addict are not when engaged in the behaviours characteristic of these conditions,

There is then no inconsistency between this theory of constraints by subpersonal modules as applied to GTS patients and addicts, and our typical intuitions of guilt attribution to psychopaths. Nevertheless, such intuitions are not in practice yet based upon a grasp of neurocognitive theory. What is the actual basis for these responses? are they justified and useful? Do they add to or distort reasoned responses based upon the practical considerations of PPDR?

Membership in the moral community

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 that, on the basis of the postulate that “ought implies can”, appears to invalidate the attribution of guilt. In an influential article, Peter Strawson (1962) 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. Such attitudes signal a subject's engagement and participation in a system of personal and social relations. Strawson grants that these attitudes 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 in the first person: subjectively, we are indeed forced to be free. We can regard Strawson as *extending this observation about the first person to the second person.* Given an essential proviso, it is not possible to regard one's interlocutor as a thing. The 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". The "social contract" is, of course, a fiction. But what underwrites the fiction is a first person acceptance of responsibility, of the *possibility of fault or guilt*. In participating in social relations with others, we effectively claim the right to be found guilty. In this way, we take upon ourselves the freedom that we are "forced" to exercise.

That, we suggest, is what accounts for 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.

There is much to recommend this view of the moral community. It does not, however, account for our intuition about the psychopath. If anyone fails to accept responsibility and the right to be found guilty it is the psychopath, yet he evokes the strongest of reactive attitudes.

This returns us to the case of Jane. We find Jane's reactions easily understandable. Yet on Strawson's position, isn't Joe, as a psychopath, an inappropriate target for such reactive attitudes? Jane expected that the court would inflict the maximal punishment for his lack of remorse. But shouldn't the court be sentencing him on the basis of *PPDR* rather than moral outrage? On the other hand, wouldn't we want the court to be applying *PPDR* principles even if Joe was not a psychopath? In fact our court system is organized in many ways to encourage the judge and jury to assume the "objective attitude"—e.g. by excluding judges or jury members

with personal ties to this case or even similar cases: in a case like this, victims of violent crimes are likely to be disqualified from the jury. Then isn't it incoherent to allow victim impact statements at all? Are the judge and jury so devoid of empathy that they don't get that Bill's death mattered to Jane? For that matter, should the sentence be any lighter if Bill didn't matter to anyone?

In short, using Strawson's moral community concept so as to exclude psychopaths is not especially helpful in understanding what is happening here. The trial is in fact a hybrid procedure that includes, on the one hand, agents of society that best serve their function by assuming objective attitudes—whether or not the criminal is a psychopath—and, on the other hand, victims with reactive attitudes for which the trial provides a forum for their public expression—again whether the criminal is a psychopath or not.

There are cases where we naturally exempt persons from usual reactive attitudes. Why not Joe? Strawson says our reactive attitudes are rooted in "…the very great importance that we attach to the attitudes and intentions toward us of other human beings…whether the actions of other people… reflect attitudes toward us of goodwill, affection or esteem…or contempt, indifference or malevolence…" We would suggest that the psychopath appropriately triggers reactive attitudes precisely because he does take a stance toward us in spite of any limitations he has—albeit one characterized by some blend of contempt, indifference and malevolence. This reflects the narcissistic dimension included in most theoretical models of the condition.

To take a stance toward others is what places us in the moral community. The psychopath does this and hence is a recipient of reactive attitudes. He is in the game. He may play it nastily or poorly—but he's in the game. By being able to take a stance we are condemned to be part of the moral community and subject to reactive attitudes, just as we are condemned to

make choices. Furthermore, whether individual acts appropriately evoke reactive attitudes depends heavily on whether they follow from integrated stances toward us. Thus our reaction to a person with Tourette's syndrome who insults us because he thinks we deserve it will be quite different than if his utterance is a verbal tic. For the psychopath to take a stance toward another requires the operation of the central processing capacities. Subpersonal modules can process and react to module-specific stimuli, but cannot take a stance, an integrated view of one person by another. To view another as taking a stance toward us amounts to recognizing the other's behaviour as proceeding from the activity of their central processing capacity.

The emergence of axiology

Before further elaborating on what is implied by taking a stance toward another, let us return to Jane. She is sophisticated enough to know that Joe would not care about her victim statement. He would not feel guilty or suffer nearly enough from the punishment he faces. Yet it was vital for Jane to speak and have someone acknowledge what Bill meant to her. Part of this was a vicarious concern for Bill's reputation. But it became increasingly clear during her treatment that the bigger fear was that no one would understand her loss—that her life and happiness for twelve years had been built around loving this man, and that he had been a worthy object of that devotion. This was where her real anxiety resided. Why?

For all of her achievements in life, Jane had only become truly fulfilled when Bill became the focus of her life. But many of her friends had been critical of her choice. He was older, had been her boss when they got involved. He had never been married and most people found him opinionated, cantankerous, and brusque. So she had an especially strong need to now have her emotional investment in him validated. Such a need is not unique. Many grieving persons describe their deep sense of aloneness and even betrayal, when they see everything going on as

usual while their world is destroyed. The sun is shining; people are laughing and active, as if their loss didn't count.

When Strawson talks of how very much it matters what stance others take toward us, this includes not just judgments about us, but importantly judgments about the validity and worth of what we live for: the values and goals around which we organize our life, our personal conception of *eudemonia*. For good reasons, we all are anxious and insecure about this.

The reason for this is that the profound locus of emergence is not primarily the moral community. Rather the community itself emerges from the truly novel development of goals and values that go beyond survival and absence of physical pain. These goals and values consist in the desire for certain consciously felt qualities linked to certain conditions in the outside world and our relationship to it. This quest, the search for *eudaimonia*, represents a radical emergent shift, an axiological revolution. Something like that is detectable even in other species. In fact it is easy to see a smooth move from mere mechanical tropisms to conscious states of positive and negative valence that modulate adaptive behaviour. But in other animals such conscious states can be seen as mere signals to act. They do not require self-consciousness or reflection. The unchallenged goal is still biologic success—of the individual organism and/or the genes. But with the eudaimonistic emergence—an event that may be intimately linked to the emergence of linguistic capacity—an organism that is organized around these biologic ends is essentially hijacked for another sort of goal altogether: the creating and perpetuating of satisfying self-conscious states, desired for their own sake, linked with an awareness of conditions in the world.

We have considerable choice in what eudaimonistic goals we create, within the constraints imposed by our biology. Hence considerable variation among goals pursued will appear from person to person. The danger here is that this interpersonal variation and the

important subjective aspect of our goals can lead us to an axiological solipsism: that our goals are just events in our own minds, having no relevance or meaning in the larger world. With respect to the non-human world, in fact they don't.

We often talk about emergence misleadingly as if it were a conflict-free flowering of levels of complexity out of simpler systems. But many of the problematic and profound instances of emergence can be conceptualized as a theft or hijacking of an existing system, organized around one set of ends, in the service of a novel sort of goal that often flies in the face of the pre-existing ones that also continue to operate. Thus biological systems perpetuate themselves and replicate as systems far removed from thermodynamic equilibrium in defiance of the pull of entropy, the natural direction of the physical-chemical world from which they emerge. Death shows they can't win; reproduction shows they don't completely lose. The Darwinian mechanisms of biological systems care nothing for our happiness; but they are hijacked to pursue new sorts of goals in eudaimonistic systems. It is interesting in this regard to note the central role theft has played in many religious mythologies explaining the origin of the human condition – e.g. Eve's theft of the forbidden fruit, Prometheus' theft of fire. The pre-existing natural order is violated, and human nature is never the same. We can understand this clearly in terms of the difference between the Aristotelian concept of final causes and the modern, "aetiological" concept described above. Aristotle's concept of nature implied that despite their essential "rational" differences, human *eudaimonia* was of the same basic kind as that of other living things: a matter of thriving based on conforming to one's essential nature. On the modern conception, our emergent goals are individual and social, elaborated in conversation, deliberation, debate and sometimes inner conflict (de Sousa 2007).

If nature at large cannot validate our goals, and unless we personalize the cosmos as a goal-driven god, we can avoid solipsism only by appealing to our fellow humans. This is precarious, for others see the world differently, especially in modern, less traditional societies. And what we want from others is not merely an acknowledgment of our formal right to have and pursue our own goals, but an assurance that our pursuits are seen by others as worthy. For with the rise of the axiological realm, we can fail in two ways: not achieving our goals, and having goals unworthy of effort. Others may judge our goals as trivial, idiosyncratic, unworthy, misguided, illusory or delusional. We need to know how they "play' in the larger world. We know we are too prone to self-deception to trust our own assessments alone.

Hence one measure of the quality of our goals is that they make sense to some community of reasonable others. The support of and participation in such a community is of profound value. This does not mean blind acceptance of every goal of another as worthwhile. Part of our job is to help others progressively modify their ends in constructive ways. But this very capacity to take some basic integrated stance toward others and their ends thrusts us—ready or not—into a moral community and makes us subject to the reactive attitudes. This is true whether or not we are willing and/or able to embrace a "mutuality of guilt and responsibility."

The Moral Community and the Axiological Realm

The heavy reliance on others to support our personal axiology begins in childhood. We need others to help us successfully form our own ideas of what is worthwhile. Many persons find themselves in a psychiatrist's office to deal with derailments of this process. Irene is not an unusual patient in this regard. Raised by alcoholic and narcissistic parents, then married for years to a successful psychopathic businessman, she always saw her worth in focusing on the goals of others and was exploited in all her relationships. After years of therapy she left her husband and

divested from many of her acquaintances; she has become determined to focus on her own fulfillment and on developing more reciprocal relationships. But at age fifty-nine, in spite of all the growth, and financial resources that open many options, she finds herself unable to even imagine what would constitute fulfillment for her. The resulting paralysis is the focus of our current work. Irene had never received the most basic validation growing up for the framing and pursuit of her own axiology. Learning that now is a daunting task

Throughout our lives we continue to rely on the stance that key others take toward our axiological projects. In a marriage, for instance, we want more than that our partner acknowledge our right to pursue our life's work. We want them to embrace the worthiness of that pursuit, to integrate it with those ends that make life worthwhile for them. Individuals who share interests are likely to form associations in the community not only to develop those interests more effectively, but to provide mutual validation that investing energy and time in such activities is reasonable and worthwhile.

All this is to suggest that the moral community must function at two levels. At the first level, it must protect the right of individuals to flourish. This includes basic protection from unnecessary pain and suffering, as well as provision of a space for each individual to develop and pursue her own chosen projects. At the level of law, a system of rules that require respect for the rights of others to flourish may suffice. And *PPDR* may give enough guidance as to how best to enforce it. However, it can be argued that the moral community needs to take account of the fact that a human nervous system is not a simple deterministic system, but a chaotic one. (Heinrichs, 2006) This may preclude sufficient predictability for objective decisions based on *PPDR* to guide individual decisions without some appeal to our reactive attitudes.

On the basis of the foregoing analysis, the moral community plays a crucial role in enhancing the individual's successful creation of an integrated notion of her own flourishing. This need not constrain the vast variety of individual goals. For while the range of eudaimonistic ends that can be created is wide, it is not infinite. Biology and culture constrain what is possible and what is prudent. Thus it is reasonable to expect a *techne* of raising children to flourish and a therapeutic *techne* when the former goes awry. Some social practices and institutions will be more helpful in enhancing these ends than others. The seeming incoherence of permitting victim statements at a trial becomes understandable in this light. While a criminal trial generally takes the objective stance and applies *PPDR* to address the first level of formal rights to flourish, victim statements make room for a person to seek validation of the meaning of their loss from the community at the second level. Judged by its overall performance at supporting this second level, modern Western democracies appear far less impressive. And this leads us back to Aristotle, who emphasized society's mandate to help us discover what it means to lead a good and full human life. And we believe that a careful analysis of what neuroscience and psychology are discovering can tell us enough to about what makes a good or bad candidate for a eudaimonistic end, to make this more than a vacuous task.

References

Alexander, B. K. (1990). *Peaceful measures: Canada's way out of the "War on Drugs"*. Toronto: University of Toronto Press

Allen, Colin, Marc Bekoff, and George Lauder, eds. 1998. *Nature's Purposes: Analyses of Function and Design in Biology*. Cambridge, MA: MIT Press, A Bradford Book.

Augustine. 1909-14. *Connfessions*. Trans. Edward Pusey. Harvard Classics. New York: Collier and Son.

Barrett, Justin L. 2004. *Why Would Anyone Believe in God?* Walnut Creek, AltaMira.

Berridge, Kent C., and Elliot.S. Valenstein. 1991. "What Psychological Process Mediates Feeding Evoked by Electrical-Stimulation of the Lateral Hypothalamus." *Behavioural Neuroscience* 105(1, February): 3-14.

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason and the Human Brain*. New York: G. P. Putnam's Sons.

de Sousa, Ronald. 1999. "Groupies: Review of E. Sober and D. Wilson, *Unto Others." Semiotic Review of Books* 10(2): online at http://www.chass.utoronto.ca/epc/srb/srb/groupies.html.

de Sousa, Ronald. 2007. *Why Think? Evolution and the Rational Mind.* New York: Oxford University Press.

De Winter, Willem. 1997. "The Beanbag Genetics Controversy: Towards a Synthesis of Opposing Views of Natural Selection." *Biology and Philosophy* 12(2, February): 149-84.

Descartes, René. [1649] 1989. *The Passions of the Soul*. Indianapolis: Hackett.

Dutton, D. G., and A. P. Aron. 1974. "Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety." *Journal of Personality and Social Psychology* 30:510-17.

Fodor, Jerry. 1983. *Modularity of Mind.* Cambridge, MA: MIT Press

Gardner, Eliot L., and James David. 1999. "The Neurobiology of Chemical Addition." In *Getting Hooked: Rationality and Addiction*, ed. Jon Elster and Ole-Jørgen Skog. Cambridge, New York, Melbourne: Cambridge University Press.

Gavrilets, Sergey. 2004. *Fitness Landscapes and the Origin of Species*. Princeton, Oxford: Princeton University Press.

Griffiths, Paul E., and Russell D. Gray. 1994. "Developmental Systems and Evolutionary Explanations." *Journal of Philosophy* 91:277-304.

Haggard, Patrick. 2005. "Conscious Intention and Motor Cognition." *Trends in Cognitive Sciences* 9(6): 290-95.

Hare, Robert.D, and Paul Babiak. 2006. *Snakes in Suits*. New York: Harper Collins.

Heinrichs, Douglas. 2006. "Antidepressants and the Chaotic Brain: Implications for the Respectful Treatment of Selves." *Philosophy, Psychiatry, and Psychology* 12,3: 215-27.

Hirstein, William. 2005. *Brain Fiction: Self-Deception and the Riddle of Confabulation*. Cambridge, MA: MIT.

Hoffman, Donald D. 1998. *Visual Intelligence: How we Create What we See*. New York: Norton.

Hofstadter, Douglas R. 1980. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Random House, New York.

Hurka, Thomas. 2000. *Virtue, Vice and Value*. Oxford; New York: Oxford University Press.

Kalanithi, Paul, Wei Zheng, Yuko Kataoka, and et al. 2005. "Altered Parvalbumin-Positive Neuron Distribution in Basal Ganglia of Individuals with Tourette Syndrome." *Proceedings of the American Academy of Sciences* 102(37, 13 September): 13307-12.

Kant, Immanuel. [1785] 1959. *Fundamental Principles of the Metaphysics of Morals*. Trans. Lewis White Beck. The Library of Liberal Arts. Indianpolis: Bobbs-Merrill.

LaLumiere, Ryan and Peter Kalivas. 2008. "Cocaine Addiction: Mechanisms of Action." *Psychiatric Annals* 38:252-8.

Libet, Benjamin. 1985. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioural and Brain Sciences* 8:529-66.

Maynard Smith, John. 1984. "Game Theory and the Evolution of Behaviour." *The Behavioural and Brain Sciences* 7:95-126.

Maynard Smith, John, and Eörs Szathmáry. 1995. *The Major Transitions of Evolution*. Oxford; New York: W.H. Freeman.

Millikan, Ruth. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press, A Bradford Book.

Murphy, Dominic. 2006. *Psychiatry in the Scientific Image*. Cambridge, MA: MIT Press

Oyama, Susan. [1985] 2000. *The Ontogeny of Information: Developmental Systems and Evolution*. Second Edition revised and expanded. Durham, NC: Duke University Press.

Proust, Joelle. 2005. *La Nature de la Volonté*. Le temps d'une question. Paris: Gallimard, folio essais.

Schroeder, Tim. 2005. Moral responsibility and Tourette Syndrome. *Philosophy and Phenomenological Research*, 71/1:106-123.

Sober, Elliott, and David Sloan Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behaviour*. Cambridge, MA: Harvard University Press.

Sommerhoff, Gerd. 1950. *Analytical Biology*. Oxford: Oxford University Press.

Strawson, P. F. 1962. *Freedom and Resentment.* Annual philosophical lecture, Henriette Hertz trust 1962. London: Oxford University Press.

Taylor, Charles. 1964. *The Explanation of Behaviour*. International Library of Philosophical and Scientific Method. London: Routledge and Kegan Paul.

Vignemont, Frederique de, and Tania Singer. 2006. "The Empathic Brain: How, When and Why?" *Trends in Cognitive Science* 30(10): 435-44.

Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. Cambridge MA: MIT Press.

Wieseltier, Leon. 2006. "The God Genome." *New York Times*, 2006, 19 February.

Wilson, Edward O. 1998. *Consilience*. New York: Knopf.

Wilson, Timothy D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA; London: Harvard University Press, Belnap.