



## Feeling Reasons

Patricia Smith Churchland  
Philosophy Department and Salk Institute  
UCSD, La Jolla California

### 1. Introduction - The Social Significance of Agent Autonomy and Responsibility

Much of human social life depends on the notion that agents have control over their actions and are responsible for their choices. We assume that it is sensible to punish and reward behavior so long as the person was in control and chose knowingly and intentionally. Without the assumptions of agent control and responsibility, human social commerce is hardly conceivable. As members of a social species, we recognize co-operation, loyalty, and reciprocation as prominent features of the social environment, and we react with hostility when group members disappoint socially salient expectations. Inflicting disutilities on the socially renegade and rewarding civic virtue helps restore the standards. In other social species too, social unreliability, such as failures to reciprocate grooming or food-sharing, provoke a reaction likely to cost the renegade animal or his kin, sooner or later. For example, de Waal (1982) observed that chimpanzees that renege on a supportive coalition when loyalty is needed will later suffer retaliation. In social mammals at least, mechanisms for keeping the social order seems to be part of what evolution bequeathed to brain circuitry (Clutton-Brock and Parker 1995). The stability of the social-expectation baseline is sufficiently important to survival that individuals are prepared to incur some cost in enforcing those expectations. Just as an anubis baboon learns that tasty scorpions are to be found under rocks but cannot just be picked up, so it learns that failure to reciprocate grooming when it is duly expected may yield a smart slap. Much of behavior is guided by the expectation of certain consequences -- not only what will happen in the physical world, but including also what will happen in the social world (Cheney and Seyfarth 1990; de Waal 1989).

What is it -- for us or baboons or chimpanzees -- to have control over one's behavior? Are we really responsible for our choices and decisions? Will neuroscientific understanding of the neuronal mechanisms for decision-making change how we think about these fundamental features of social commerce? These are some of the

questions I wish to consider in this essay.

## **2. Are we responsible and in control if our choices and actions are caused?**

A venerable tradition bases the conditions for free will and control on a contrast between being caused to do something and not being so caused. For example, if someone pushes me from behind and I bump into you, then my bumping you was caused by the push; I did not choose to bump you. Examples conforming to this prototype have given credence to the idea that in order for a choice to be free, it must be uncaused. That is, it is supposed that a free choice is made when, without prior cause and without prior constraints, a decision comes into being and an action results. This *contracausal* construal free choice is known as libertarianism. (See C. A. Campbell 1957) Is it plausible?

As Hume pointed out in 1739, the answer is surely no. Hume argued that our choices and decisions are in fact caused by other events in the mind -- desires, beliefs, preferences, feelings, and so forth. Nor need the precipitating events, whether described as mental or as neuronal, be conscious. He also made the much deeper and more penetrating observation that agents are not considered responsible for their choices made *unless* they are caused by our desires, intentions, and so forth. Randomness, pure chance, utter unpredictability, are not preconditions for control. Hume puts the issue with memorable compactness:

where [actions] proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honor if good, nor infamy, if evil. (p. 411)

Logic reveals, Hume argued, that responsibility is actually inconsistent with libertarianism (uncaused choice). Someone may choose to climb onto his roof because he does not want the rain to come in his house, he wants to fix the loose shingles, and he believes that he needs to get up on the roof to do that. His desires, intentions and beliefs are part of the causal antecedents resulting in his choice. If, without determining desire and belief, he simply went up onto the roof -- as it were, for no reason -- his sanity and hence his control is seriously in doubt. More generally, a choice undetermined by anything the agent believes, intends or desires is actually the kind of thing we consider out of the agent's control and not the sort of thing for which we hold someone responsible. Furthermore, desires or beliefs, were they uncaused rather than caused by other stable features of the person's character and temperament, are likewise inappropriate preconditions for responsible choice. (See also Hobart 1934.)

Neither Hume's argument that choices are internally caused nor his argument showing that libertarianism is absurd has ever been

convincingly refuted. (For disagreements with Hume, see Kenny 1989) Notice, moreover, that his arguments hold whether or not one thinks of the mind as a separate Cartesian substance, or as pattern of activity of the physical brain; whether one thinks of the etiologically relevant states as conscious or unconscious. If anything in philosophy could count as a result, Hume's argument on free will does. Nonetheless, the idea that randomness in the physical world is somehow the key to what makes free choice free remains appealing to those inclined to believe that free choice must be uncaused choice. The appeal of quantum mechanics, chaos, and so forth, as a "solution" to the problem of free will and responsibility generally derives from an intuition innocent of exposure to Hume's result.

If all behavior is caused, what is the difference between voluntary and involuntary actions? When, if ever, is an agent responsible? Many possibilities have been explored in attempts to explain how the notions of control and responsibility can make sense in the context of causation -- of determinism. To begin with, it is clear that the distinction between internal and external causes will not suffice to distinguish the voluntary from the involuntary. A patient with Huntington's disease cannot prevent himself from making choreoform movements; a sleepwalker may unplug the phone or kick the dog; a phobic patient may have an overwhelming urge to wash her hands. The cause of the behavior in each of these cases is internal -- in the subject's brain. Yet the behavior is considered to be out of the agent's control.

Another strategy is to base the distinction on felt differences in inner experience between those actions we choose to do, and those over which we feel we have no control. Thus it allegedly feels different when we evince a cry as a startle response to a mouse leaping out of the compost heap, and when we cry out to get someone's attention. Is introspection a reliable guide to responsibility? Can introspection distinguish those internal causes for which we are responsible from those for which we are not? (See also Crick 1994) Probably not. There are undoubtedly many cases where introspection is no guide at all. Phobic patients, obsessive-compulsive patients, and Touretters are obvious examples that muddy the waters. The various kind of addictions present further difficulties. A contented smoker typically feels that the desire for a cigarette is indeed his, and that his reaching for a cigarette feels as free as reaching to turn on the television. Not so the smoker who is trying to quit the habit. The increase in intensity of sexual interest and desire at puberty is surely the result of hormonal changes on the brain, not something over which one has much control. Yet engaging in certain activities such as ogling the opposite sex, feels as free as tying one's shoes. More problematic perhaps, are the many examples from everyday life where one may suppose the decision was entirely one's own, only to discover that subtle manipulation of desires had in fact been the decisive factor. According to the fashion standards of the day, one finds certain clothes beautiful, others frumpy, and the choice of wardrobe seems, introspectively, as free as any choice. There is no escaping, however, the fact

that what is in fashion has a huge effect on what we find beautiful in clothes, and this affects not only choice in clothes, but also such things as aesthetic judgment regarding plumpness or slenderness of the female body.

Social psychologists have produced dozens of examples that further muddy the waters, but a simple one will convey the point. On a table in a shopping mall, experimenters placed ten pairs of identical panty hose, and asked shoppers to select a pair, and then briefly explain their choice. After selection, the choosers referred to color, denier, sheerness and so forth, as their rationale. In fact, there was a huge position effect: shoppers tended to pick the pantyhose in the rightmost position on the table, and of course in fact the panty hose were identical and differed not at all in color, sheerness and so on. None of the subjects considered position to be a factor, none of them referred to it as a basis for choice, yet it clearly was so. Other examples of priming, subliminal perception, emotional manipulation, and so forth make introspection a highly unreliable guide.

In a different attack on the problem, philosophers have explored the idea that if the choice was free, the agent *could have chosen otherwise*, that in some sense, the agent had the power to do something else. (See Taylor 1974; Kenny 1989) The weakness in the strategy shows up when we ask further, "what exactly does *that* mean? " If all behavior has antecedent causes, then "could have done otherwise" seems to boil down to "*would have done otherwise if antecedent conditions had been different*". Accepting that equivalence means the criterion is too weak to distinguish between the shouted insults of a Touretter whose tics include random and undirected outbursts ("idiot, idiot, idiot") and those of a member of parliament responding to an honorable member's proposal ("idiot, idiot, idiot"). In both cases, had the antecedent conditions been different, obviously the results would have been different. Nevertheless, we hold the parliamentarian responsible, but not the Touretter.

In our legal as well as our daily practice, the pattern is to accept certain prototypical conditions as excusing a person from responsibility, but to assume him responsible unless specific exculpatory condition obtains. In other words, responsibility is the default condition; excuse from and mitigation of responsibility has to be established. The set of conditions regarded as exculpatory can be modified as we learn more about behavior and its etiology. Thus a child Touretter might have been smacked for his ticcig outbursts in public, until it is understood that the tics are not within his control and that punishment is totally inefficacious.

Aristotle was the first to articulate this strategy, and the core of his ideas on this matter is still reflected in much of human practice, including current legal practice. In his systematic and profoundly sensible way, Aristotle pointed out that it is a necessary condition that the cause be internal to the agent, but in addition, he characterized as involuntary, actions produced by

coercion and actions produced in certain kinds of ignorance. As Aristotle well knew, however, no simple rule demarcates cases here. Clearly, some ignorance is not considered excusable, when it may be fairly judged that the agent *should* have known. Additionally, in some cases of coercion, the agent is expected to resist the pressure, given the nature of the situation. As Aristotle illustrates in his own discussion of such complexities, we seem to proceed to deal with these cases by judging their similarity to uncontroversial and well-worn prototypes. This run-of-the-mill cognitive strategy is reflected in the fundamental role that precedent law is accorded in determining subsequent judgments. (See more extended explanations in P. M. Churchland 1995)

It is unlikely that there exists a sharp distinction between the voluntary and the involuntary -- between being in control and being out of control -- either in terms of behavioral conditions, or in terms of the underlying neurobiology. The differences are differences in degree, not a clean bifurcation specified by necessary and sufficient conditions. An agent's decision to change television channels may be more in his control than his decision to pay for his child's college tuition, which may be more in his control than his decision to marry his wife, which may be more in his control than his decision to turn off the alarm clock. Some desires or fears may be very powerful, others less so, and we may have more self-control in some circumstances than in others. Hormonal changes, for example in puberty, make certain behavior patterns highly likely, and in general, the neurochemical milieu can have a powerful effect of the strength of desires, urges, drives and feelings.

As I shall suggest in Section 4 however, at opposite ends of the self-control spectrum are prototypical cases that differ sufficiently in behavioral and internal features to provide a foundation for a basic, if somewhat rough-hewn, distinction between being in control and not; between being responsible and not. It will also be apparent as we reflect on the spectrum's end points that there are *many* parameters relevant to being in control. In our current state of neurobiological and behavioral knowledge, we do not know how to specify all those parameters. Nevertheless, we do know now that activity patterns in certain brain structures, including the amygdala, hypothalamus, somatosensory cortices, and ventromedial frontal cortex are important, and that levels of alleged neuromodulators, such as serotonin, dopamine, and norepinephrine, as well as hormones, play a critical role. Ultimately, as shall explore later, a range of optimal values may be specifiable, and therewith, a contrasting range that is clearly sub optimal.

### **3. Are we more in control and more responsible to the degree that emotions plays a lesser role and reason plays a greater role?**

A view with deep historical roots assumes that in matters of practical decision, reason and emotion are in opposition. To be in

control, on this view, is to be maximally rational. To that end, one must maximize suppression of emotions, feelings, and inclinations. Emotion is considered the enemy of morality, and consequently moral judgment must be based on reason detached from emotion.

Immanuel Kant is the philosopher best known for adopting something rather close to this view. In his moral philosophy, Kant saw human agents as attaining virtue only as they succeed in downplaying feeling and inclination and giving reason complete control. He says: "The rule and direction for knowing how you go about [making a decision], without becoming unworthy of it, lies entirely in your reason." (From *Fragments of a Moral Catechism*.) In Kant's view, we would be perfectly rational save for the inclinations, feelings, and desires based in our bodies. The perfect moral agent, according to Kant, is one whose decisions are perfectly rational and are detached entirely from emotion and feeling. (Ronald de Sousa calls such an agent a "Kantian monster" See de Sousa 1990, p. 14) The kinds of cases that inspire Kant's veneration of reason and his suspicion of the passions are the familiar "heart-over-head" blunders where the impassioned do-gooder makes things worse, when the long term consequences were neglected while immediate need was responded to, when the fool does not look before he leaps. A powerful rationalist assumption underlies the great bulk of ethical theory in the latter half of this century. For example, the Kantian framework permeates the work of Nagel (1970), Rawls (1971), Gewirth (1978) and Donagan (1977).

Understanding the consequences of a plan, both its long and short term consequences, is obviously important, but is Kant right in assuming that feeling is the enemy of virtue, that moral education requires learning to disregard the bidding of inclination? Would we be more virtuous, or more educable morally, were we without passions, feelings, and inclinations? Not according to David Hume, against whom Kant was probably reacting. Hume asserted that "...reason alone can never be a motive to any action of the will; and secondly, it can never oppose passion in the direction of the will." (p 413) As he later explains: "T'is from the prospect of pain or pleasure that the aversion or propensity arises towards any object: And these emotions extend themselves to the causes and effects of that object, as they are pointed out to us by reason and experience." (p. 414) As Hume understands it, reason is responsible for delineating the various consequences of a plan, and thus reason and imagination work together to anticipate problems and payoffs. But feelings, informed by experience, are generated by the mind-brain in response to anticipations, and incline an agent towards or against a plan.

Common culture also finds something not quite right in the image of nonfeeling, nonemotional rationality. In the highly popular television series, *Star Trek*, three of the main characters are portrayed as capable of varying degrees of emotional response. The pointy-eared semi-alien, Mr. Spock, is typified by the absence of emotion. In trying circumstances, his head is cool, his approach

is calm. He faces catastrophe and narrow escape with easy-handed equanimity. He is puzzled by the humans' propensity to anger, fear, love and sorrow, and correspondingly fails to predict their appearance. Interestingly, Mr. Spock's cold reason sometimes results in bizarre decisions, even if they have a curious kind of 'logic' to them. By contrast, Dr. McCoy is found closer to the other end of the spectrum. Individual human suffering inspires him to risk much, to ignore future costs, or to fly off the handle, often to Mr. Spock's taciturn evaluation, "but that's illogical". The balance between reason and emotion is more nearly epitomized by the legendary and beloved Captain Kirk. By and large, his judgment is wise. He can make tough decisions when necessary, he can be merciful or courageous or angry, when appropriate. He is more nearly the ideal Aristotle identifies as the practically wise man.

Neuropsychological studies are highly pertinent to the question of the significance of feeling in wise decision-making. Research by the Damasio and their colleagues on a number of patients with brain damage shows that when deliberation is cut off from feelings, decisions are likely to be poor. Consider the patient SM whose amygdala has been destroyed and who lacks normal feelings of fear. She does not process fear signals normally, she does not recognize feelings of fear in herself, and does not evince normal facial expressions in fearful circumstances. SM does have a concept of fear of sorts, and she can tell when a human face shows a fear response. In complex circumstances, with no access to gut feelings of unease and fear, she is as likely as not to make a decision that normally wired people could easily foresee to be contrary to her interest. Whereas a normal subject would say he has an uneasy feeling about someone who is in fact predatorial, SM generates no such feelings. In a rather more complex way, the point is dramatically illustrated by the patient, EVR who first came to the Damasio's lab at the University of Iowa College of Medicine more than a decade ago.

A brain tumor in the ventromedial region of EVR's frontal lobes had been surgically removed earlier from EVR, leaving him with bilateral lesions. Following his surgery EVR enjoyed good recovery and seemed very normal, at least superficially. For example, he scored as well on standard IQ tests as he had before the surgery (about 140). He was knowledgeable, answered questions appropriately, and so far as mentation was concerned, seemed unscathed by his loss of brain tissue. EVR himself voiced no complaints. In his day-to-day life, however, a very different picture began to emerge. Once a steady, resourceful and efficient accountant, now he made a mess of his tasks, came in late, failed to finish easy jobs, and so forth. Once a reliable and loving family man, his personal life became a shambles. Because he scored well on IQ tests and because he was knowledgeable and bright, EVR's problems seemed to his physician more likely to be psychiatric than neurological. As we now know, this diagnosis turned out to be entirely wrong.

The case of EVR is by no means unique, and there are a number of

patients with a similar lesion and a comparable behavioral profile. After studying EVR for some time, and comparing him to other cases of similar damage, the Damasio and their colleagues began to devise new tests to determine what about EVR's emotional responses were not in the normal range. For example, shown horrifying or disgusting pictures, his galvanic skin response (GSR) was flat. Normals, in contrast, show a huge response while viewing such pictures. On the other hand, he could feel fear or pleasure in uncomplicated, more basic, situations. During the following years, new and more revealing tests were devised to try and probe more precisely the relation between reasoning logically on the one hand, and acting in accordance with reason on the other. For there was no doubt that EVR could evince the correct answer to questions concerning what would be the best action to take (e.g. defer a small gratification now for a larger reward later), but his own behavior often conflicted with his stated convictions (e.g. he would seize the small reward now, missing out on the large reward later) (Saver and Damasio 1991).

Antoine Bechara, working with the Damasio, developed a particularly revealing test. In this test, a subject is presented with four decks of cards, and told that his task is to make as much profit as possible, given an initial loan of money. Subjects are told to turn over cards, one at a time, from any of the four decks. They are not told how many cards can be played (a series of 100) or what the payoffs are from any deck. One has to discover everything by trial and error. After turning over each card, the subjects are rewarded with an amount of money, and on some cards, they may also be penalized and be required to pay out money. Behind the scenes, the experimenter designates two decks, C and D, to be low-paying (\$50) and to contain some moderate penalty cards; two other decks, A and B, pay large amounts (\$100) but contain very high penalty cards. Things are rigged so that players incur a net loss if they play mostly A and B, but do well if they play mostly C and D decks. Subjects cannot calculate exactly losses and gains because there is too much mentally to keep track of. Instead, subjects must generate a sense of what strategy will work to their advantage.

During the game, normal controls come fairly quickly to stick mainly with low-paying, low-penalty decks (C and D) and make a profit. What is striking is that subjects such as EVR (ventromedial frontal damage), tend to end with a loss because they choose mainly high-paying decks despite the profit-eating penalty cards. Subjects with brain damage to regions other than ventromedial behave like controls. As Bechara et al. note, even after repeated testing on the task as long after as a month or as short after as twenty-four hours, EVR continued to play heavily the losing decks. When queried at the trial end, he verbally reports that A and B are losing decks. To put it rather paradoxically, *rationaly* EVR does indeed know what the best long-run strategy is, but in exercising choice in actual action, he goes for short run gain, incurring long run loss. Is EVR merely showing frontal perseveration? No, because he does score normally on the Wisconsin card-sorting task, in contrast to perseverative

patients. In any case, he does sometimes try other decks. To make matters more difficult for the Kantian ideal, his judgments of recency and frequency are flawless, his knowledge base and short term memory are intact. (Bechara et al. 1994) Moreover, EVR can articulate well enough the future consequences of alternative actions, so the problem cannot be lack of understanding of what might happen. In sum, what seems chiefly to be amiss here is not EVR's capacity to reason; rather, it is the inability of emotions to affect his reason and decision-making.

That his "pure reasoning", displayed verbally, and his "practical decision-making", displayed in choice, were so at odds suggested to the Damasio that the real problem lay with EVR's lack of emotional responsivity to situations that involved some understanding of the meaning and implications of the events. This is consistent with his lack of skin response while viewing emotionally charged photographs. That is, although EVR would react normally to simple conditions such as a loud noise or a threat of attack, he failed to respond in more complex situations whose significance might involve more subtle or culturally-mediated features, such as the social consequences of failure to complete jobs or the future consequences of a sudden marriage to a prostitute or profit-making in the Bechara gambling task. EVR and similar subjects seem to have a kind of insensitivity to the significance of future consequences, whether they are rewarding, as in the gambling task, or whether they are punishing, as in a reverse gambling task. This insensitivity seems best understood in terms of the failure to provide a "value mark", via somatic states, to the various options. (Bechara et al. 1994; A. Damasio 1994; Adolphs, this volume, A. Damasio, this volume.)

Further results came from the analysis of skin conductance data taken by a galvanometer placed on the arm of each subject during the gambling task. (Damasio et al. in press) In the gambling task, neither controls nor frontal patients showed a skin response to card selections in the first few plays of the game (selections 1-10). By about the tenth selection, however, controls began to exhibit a skin response immediately prior to beginning to reach for the "bad" decks. When queried at this stage about how they were making their choices, controls (and frontal patients) said they had no idea whatever, they were just exploring. By about selection 20, controls continued to get a consistent skin response just before reaching for the "bad" decks. In their verbal reports, controls said that they still did not know what was the best strategy, but that they had a feeling that maybe decks A and B were "funny". By selection 50, controls typically could articulate -- and follow -- the winning strategy. Frontal patients never did show a skin response in reaching for any deck. What is so striking here is that in controls, good choice was in some measure biased by the feeling even before subjects were aware of the feeling, and well before they could articulate the winning strategy. That many of our daily choices are likewise biased without our being aware of the feeling, seems altogether likely.

The significance for choice of feeling, and of unaware biasing by

feeling, has implications for the economists' favored model of "rational choice". According to this model, the ideally rational (wise) agent begins deliberation by laying out all alternatives, calculating the expected utility for each alternative based on probability of each outcome multiplied by the value of (goodies accruing to ) each outcome. He ends by choosing the alternative with the highest expected utility score. In light of the data just considered, this model seems highly unsatisfactory. At best, it probably applies to a small range of highly quantifiable problems, but even then comes into play after "cognition-cum-feeling" brings to awareness the restricted set of "feels-reasonable" alternatives. At any rate, the economists' model is unlikely to come even close to giving the whole story of rational choice.

A major idea in the Damasio's work on decision-making is that *representation of changes in body state*, comprising visceral feeling as well as musculoskeletal signals (both external 'touch' sensations and sympathetic system changes in skin), plays an essential role in biasing choice. Body-state representation systematically integrates diverse changes in information originating in the sympathetic system. As future plans and possible-plans develop, the imagination generates representations of plan sequelae, and to these, as well as to perceptually-driven representations, visceral responses are generated, via mediation of the amygdala and hypothalamus. Somewhat more elaborately, therefore, the Damasio envisage a complex to-ing and fro-ing of signals between thalamo-cortical states and changes in the body as being the crucial pathways for self-representation and for sensible decision-making.

The connection between this hypothesis and the case of EVR and others with similar lesions (ventromedial frontal) is obvious. The point is not that EVR-type patients feel nothing at all. Rather, it is that in those situations requiring imaginative elaboration of the consequences of an option, feelings are not generated in response to the imagined scenario. This is because the ventromedial frontal region needed for integration of body-state representation and fancy "scenario-spinning" is disconnected from the "gut feelings". Normally, neurons in ventromedial frontal cortex would project to and from areas such as the amygdala and hypothalamus that contain neurons carrying body-state signals. In patients with destruction of ventromedial cortex, the pathways are disrupted.

It is this set of complex responses, involving future-consequences-recognition, visceral changes, and feelings, that the Damasio see as inclining the person to one decision rather than another. That is, in the context of acquired cognitive-cum-emotional understanding about the world, neuronal activity in these pathways bias sensible decisions. Moreover, the biasing can begin before subjects are aware of it, and before they can articulate their inclinations, even though a particular decision may seem, introspectively, to be the outcome solely of conscious deliberation. Their hypothesis is that when EVR is confronted with a question ("should I finish this job or watch the football game

?", "should I choose from deck A or from deck C?"), his brain's body-state representation contains nothing about changes in the viscera, and hence he is missing important biasing clues that something is foolish or unwise or problematic. His frontal lobes, needed for a complex decision, have no access to information about the valence of a complex situation or plan or idea. Therefore, some of EVR's behavior turns out to be foolish and unreasonable. (For a much more full account, see Antonio Damasio's 1994 book, *Descartes' Error*.)

#### **4. Are there significant neurobiological differences between "in control" agents, and "out of control" agents?**

I am assuming there is a real difference in what we may loosely call "life success" between agents whose behavior is generally at the "in control" end of the spectrum -- agents typified by the fictional Captain Kirk, and on the other hand, agents whose behavior is often at the "out of control" end of the spectrum, typified by the obsessive-compulsive subject. To a first approximation, the behavior of the "in controls" is more conducive to their interests, long and short term, and they generally make sensible, reasonable, and wise decisions about both short and long term plans. Undoubtedly the relevant behavioral differences cannot be simply reduced to a formula for "reproductive fitness" or even "inclusive fitness", yet they are almost certainly deeply related to properties sensitive to natural selection. Are there likely to be prototypical neurobiological differences that correlate with these behavioral differences, vaguely specified though they are? Is it possible to provide an outline of the neurobiology prototypes at either end of our "in control/out of control" spectrum, based on the empirical data so far?

In a rather crude, semi-speculative, and somewhat metaphorical way, yes. Having made all those hedges, let me now be a bit bolder. To begin with, converging data from neuropsychological research, animal studies and anatomy implicate certain brain structures as especially crucial in emotional response. Assuming the Damasio hypothesis is largely correct, these structures are presumably essential for "in control" behavior. So far, the list of structures include the amygdala, somatosensory cortical areas I and II, the insula, hypothalamus, the anterior cingulate cortex, basal ganglia and the ventromedial frontal region of cortex. That there is more than one areas indicates that "control", in this vague sense, is likely a distributed function in which a number of structures participate. The aforementioned structures also heavily project to and from each other. We do also know that the connectivity between these regions is essential to normal functioning, as evidenced by subjects such as EVR, SM and others. Given the data, therefore, it may be expected that certain general dynamical properties of the neural networks in these regions probably characterize functioning in the normal range. How might we begin to specify the dynamical properties typical of the normal range?

The brain is a complex dynamical system. If we consider the

brain's neurons as defining axes in a multidimensional state space, neuronal activity can be represented as points in the state space, and patterns of neuronal activity as trajectories in that state space. Additionally, trajectories can be linked to make sequences of trajectories in that state space. Planning for a future contingency, for example how to portage around heavy rapids or how to convince a jury of Simpson's guilt, involves putting together a long and complex sequence of trajectories, first in imagination, and later in behavior. Given available data, it seems reasonable to assume that the "well-tempered" brain of a person who is typically wise, sensible and reasonable, embodies a kind of stable landscape in that state space. This means, among other things, that trajectories, appropriate to the sensory environment, through that state space, are highly stable to small perturbations. On the other hand, when the structures are damaged by lesions or by highly abnormal changes in neurochemical milieu, the landscape of the state space changes, and the trajectories become unstable and bizarre. To use a related analogy, the limit cycle characterizing the stable path in neuronal space can suddenly be replaced by very different, and behaviorally inappropriate, limit cycles.

Certain general features of the "neuronal landscape" may be particularly dependent on the neurochemicals of the widely projecting systems, such as those involving serotonin, dopamine, or norepinephrine. When the concentrations of these chemicals change by a certain threshold, they may change quite radically the terrain of the landscape, for example by flattening it to make trajectories more susceptible to perturbations, or digging deep grooves that are hard to get out of even when the environment calls for a different trajectory. The widely projecting systems appear to play a role in modulating the responses of neurons, for example, by changing the gain or by down-regulating responsivity to neurotransmitters such as glutamate. They appear to play a crucial role in sleep and dreaming, in attention, in mood, and they are found in brain stem structures that diffusely project to cortical and subcortical structures. We do not, however, understand in detail the interactions between these systems, nor exactly how changing their concentrations changes the neuronal landscape.

What is known at the behavioral level is that obsessive-compulsive disorder (OCD), for example, is very treatable by serotonin agonists such as fluoxetine (Prozac); patients gain control over their phobias and their compulsive behavior. It is well known that many cases of medical depression respond extremely well to drugs that enhance serotonin or norepinephrine, giving patients control over their anger and lassitude. Tourette's syndrome is much more controlled when patients are given serotonin agonists. Each of these interventions can be seen, in keeping with the favored metaphor, as establishing or re-establishing general features in neuronal landscape that make it stable against noise and perturbation, while allowing exploration and innovation without wild deviation from the stable trajectories.

Research from basic neuroscience as well as from lesions studies and scan studies will be needed to transform this speculative outline into a substantial, detailed, testable account of the general features of the landscape that are typical of "in control" subjects. These dynamical systems properties may be quite abstract, for "in control" individuals may differ in temperament and in cognitive strategies. As Aristotle might have put it, there are different ways to harmonize the soul. Nevertheless, the prediction is that at the very least, some such general features probably are specifiable. It is relatively easy to see that dynamical systems properties do distinguish between brains that perform certain tasks, such as walking well or poorly. What I am proposing here is that more abstract skills characterizable behaviorally, such as being a successful shepherd dog, or a competent lead sled dog, can also be specified in terms of dynamical systems properties, dependent as they are on neural networks and neurochemical concentrations. My hunch is that human skills in planning, preparing and co-operating, can likewise be specified. Not now, not next year, but in the fullness of time as neuroscience and experimental psychology develop and flourish.

##### **5. Learning what's rational and what's not.**

Aristotle would have us add here the point that there is an important relation between self-control and habit formation. A substantial part of learning to cope with the world, to defer gratification, to show anger and compassion appropriately, to have courage when necessary, involves acquiring appropriate decision-making habits. In the metaphor of dynamical systems, this is interpreted as sculpting the terrain of the neuronal state space so that behaviorally appropriate trajectories are well-grooved. Clearly, we have much to learn about what this consists in, both at the behavioral and at the neuronal level. We do know, however, that if an infant has damage in some of the critical regions, such as the ventromedial frontal cortex or amygdala, then typical acquisition of the right "Aristotelian" contours may be next to impossible, and more direct intervention may sometimes be necessary to achieve what normal children routinely achieve as they grow up.

The characterization of a choice or an action as 'rational' carries a strongly normative component, implying, for example, that it was in one's long range interest, or in the long range interest of some relevantly specified group, or not inconsistent with other things one believes or what is believed by "reasonable" people. Insofar, it is not sheerly descriptive, in contrast, for example, to describing the action as performed clumsily or with a hammer. Claiming an action was rational often carries the implication that the choice was conducive in some significant way to the agent's interests or well-being or to those of his family, and that it properly took into account the consequences of the action, both long and short term. (See also Johnson 1993) Thus the evaluative component. Though a brief dictionary definition can capture some salient aspects of what it means to be rational and

reasonable, it hardly does justice to the real complexity of the concept.

As children, we learn to evaluate actions as more or less rational by being exposed to prototypical examples, as well as to prototypical examples of foolish or unwise or irrational actions. Insofar as we learn by example, learning about rationality is like learning to recognize patterns in general, whether it be recognizing what is a dog, what is food, or when a person is afraid or embarrassed or weary. As Paul Churchland (1995) has argued, we also learn ethical concepts such as 'fair' and 'unfair', 'kind' and 'unkind', by being shown prototypical cases and generalizing to novel but relevantly similar situations. Now as we know, learning from examples is something *networks* do exceedingly well. Peer and parental feedback hone the pattern recognition network so that over time it comes to closely resemble the standard in the wider community. Nevertheless, as Socrates was fond of showing, articulating those standards is a hopeless task, even when a person successfully uses the expression 'rational', case by case. Making an algorithm for rational choice is almost certainly impossible. The systematic failure of AI research to discover how to program computers to conform to common sense is an indication of the profoundly nonalgorithmic nature of common sense, rationality and practical wisdom.

This is important, because most philosophers regard the evaluative dimension of ethical concepts to imply that their epistemology must be entirely different from that of descriptive concepts. What appears to be special about learning some concepts, such as 'rational', 'impractical' and 'fair', is only that the basic wiring for feeling the appropriate emotion must be intact. That is, the prototypical situation of something's being impractical or shortsighted typically arouses unpleasant feelings of dismay and concern; the prospect of something's being dangerous arouses feelings of fear, and these feelings, along with perceptual features, are probably an integral part of what is learned in perceptual pattern recognition.

Simple dangerous situations -- crossing a busy street, encountering a grizzly with cubs -- can likely be learned as dangerous without the relevant feelings. At least that is suggested by the evidence of the Damasio from their patient SM who, recall, suffered amygdala destruction and lacks normal fear processing. Although she can identify which simple situations are dangerous, this seems for her to be a purely cognitive, nonaffective judgment. Where her recognition is poor, however, is when she needs to detect the menace or hostility or pathology in complex social situations, where no simple formula for identifying danger is available. As suggested earlier, the appropriate feelings may be necessary for skilled application of a concept, if not for fairly routine applications. This is perhaps why the fictional Mr. Spock, lacking emotions as he is, is poor at predicting what will provoke strong sympathy or dread or embarrassment in humans.

Stories, both time-honored as well as those passing as local gossip, provide a basic core of scenarios where children imagine and feel, if vicariously, the results of various choices such as failing to prepare for future hard times (*The Ant and The Grasshopper*) or failing to heed warnings (*The Boy Who Cried Wolf*), of being conned by a smooth talker (*Jack and the Beanstock*), of vanity in appearance (*Narcissus*). As children, we can vividly feel and imagine the foolishness of trying to please everybody (*The Old Man and his Donkey*), of not caring to please anybody (Scrooge in Dickens' *A Christmas Carol*), and of pleasing the "wrong" people (the prodigal son). Many of the great and lasting stories, for example by Shakespeare, Ibsen, Tolstoy, Balzac, are rife with moral ambiguity, reflecting the fact that real life is rife with conflicting feeling and emotions, and that simple foolishness is far easier to avoid than great tragedy. Buridan's dithering ass was just silly; Hamlet's ambivalence and hesitation was deeply tragic and all too understandable. In the great stories is also a reminder that our choices are always made amidst a deep and unavoidable ignorance of many of the details of the future, where coping with that very uncertainty is something about which one can be more or less wise. For all decisions save the trivial ones, there is no algorithm for making a wise choice. (See again Johnson 1993; Flanagan 1991) For matters such as choosing a career or a mate, having children or not, moving to a certain place or not, deciding the guilt or innocence of a person on trial etc. -- these are usually complex constraint satisfaction problems.

As we deliberate about a choice, we are guided by -- and guide others by -- our reflection on past deeds, our recollection of pertinent stories, and our imagining the sequence of effects that would be brought about by choosing one option or another. Antonio Damasio calls the feelings generated in the imagining-deliberating context "secondary emotions" (Damasio 1994, p. 134 ff.) to indicate that they are a response not to external stimuli, but to internally generated representations and recollections. As we learn and grow up, we come to associate certain feelings with certain types of situation, and this combination can be reactivated when a similar set of conditions arises. Often a moral dilemma cannot be easily labeled, and instead we draw analogies between types of dilemmas: "this is like the time my father got lost in the blizzard and built a quinze"; "this is like the time Clarence Darrow defended a teacher's right to teach evolutionary biology", etc. Recognition of a present situation as relevantly like a certain past case has of course a cognitive dimension, but it also evokes feelings that are similar to those evoked by the past case, and this is important in aiding the cortical network to relax into a solution concerning what to do next.

## **6. What Happens to the Concept of Responsibility?**

We need now to return to the dominant background question motivating this essay. One very general conclusion is provoked by the foregoing discussion. On the whole, social groups work best when individuals are considered responsible agents, and hence as a

matter of practical life, it is probably wisest to hold mature agents responsible for their behavior and for their habits. That is, it is probably in everyone's interest if the default assumption in place is that agents have control over their actions and that in general, agents are liable to punishment and praise for their actions. This is of course a highly complex and subtle issue, but the basic idea is that *feeling* the social consequences of one's choices is a critical part of socialization -- of learning to be in the give and take of the group. (This pragmatic point of view, most closely associated with the ideas of Spinoza, can also be found in the classic essays of Hobart [1934] and Schlick [1939].) *Feeling* those consequences is necessary for contouring the state space landscape in the appropriate way, and that means *feeling* the approval and disapproval meted out. A child must learn about the physical world by interacting with it and bearing the consequences of its actions, or watching others engage the world, or hearing about how others engage the world. As with social animals generally, learning about the social world involves cognitive-affective learning, directly or indirectly, about the nature of the social consequences of a choice. This must of course be consistent with reasonably protecting the developing child, and with compassion, kindness and understanding. In short, I do not want the simplicity of the general conclusion to mask the tremendous subtleties of child-rearing. Underlying all the necessary subtlety, however, the basic pragmatist point is just this: if the only known way for "social decency" circuitry to develop requires that the subject generate the relevant feelings pursuant to social pattern recognition, and if that, in turn, requires experiences of praise (pleasure) and blame (pain) consequent upon his actions, then treating the agent as responsible for behavior is a pragmatically justified operating assumption. That is, it is justified by its practical necessity.

This of course leaves it open that under special circumstances agents should be excused from responsibility or accorded diminished responsibility. In general, the law courts are struggling, case by case, to make reasonable judgments about what those circumstances are, and no simple rule really works. Neuropsychological data are clearly relevant here, as for example in cases where the subjects' brains show an anatomical resemblance to the brain of EVR or SM. Quite as obviously, however, the data do not show that no one is ever really responsible and deserving of punishment or praise. Nor do they show that when life is hard, one is entitled to avoid responsibility.

Is direct intervention in the circuitry morally acceptable? This too is a hugely complex and infinitely ramifying issue. My personal bias is twofold: first, that in general, at any level, be it ecosystem or immune system, intervening in biology always requires great caution. When the target of the intervention is the nervous system, then caution by many more orders of magnitude is wanted. Still, not taking action is nevertheless doing something, and *acts of omission can be every bit as consequential as acts of commission*. Second, the movie, *Clockwork Orange*, immediately associated with the very idea of direct intervention in criminal

law, probably had a greater impact on our collective amygdaloid structures than it deserves to have. Certainly some kinds of direct intervention are morally objectionable. So much is easy. But *all* kinds? Even pharmacological? Is it possible that some forms of nervous system intervention might be more humane than lifelong incarceration or death? It seems to me likely that the general answer is yes. I do not know the detailed answers to these questions, but given what we now understand about the role of emotion in reason, perhaps the time has come to give them a careful, calm and thorough reconsideration. Guided by Aristotle, we may say that these are, *au fond*, pragmatic questions concerning the well-functioning of certain social animals, namely hominids.

## 7. Conclusions

I have considered three vintage philosophical theses in the context of new data from neuroscience: (1) feelings are an essential component of viable practical reasoning about what to do, (David Hume) and (2) moral agents come to be morally and practically wise not by dint of "pure cognition", but by developing through life experiences the appropriate *cognitive-affective* habits (Aristotle), and (3) agents need to acquire the cognitive-connative skills to evaluate the consequences of certain events and the price of taking risks, and hence must be treated as responsible agents (Hobart [1934], Schlick [1939]) . Each of the theses has been controversial and remains so now; each has been the target of considerable philosophical criticism. Now, however, as the data come in from neuropsychology as well as experimental psychology and basic neuroscience, the empirical probability of each seems evident. One may interpret Damasio's book, *Descartes Error*, as the beginning of a neurobiological perspective on the ideas of Aristotle and Hume. In this evolving scientific context, many important social policy questions must be considered afresh, including those concerned with the most efficacious means, consistent with other human values, for achieving civil harmony. Much, much more, of course, needs to be learned, for example about the reward circuits in the brain, about pleasure and anxiety and fear. Philosophically, the emphasis with respect to civic, personal, and intellectual virtue has been focused almost exclusively on the purely cognitive domain, with the affective domain largely left out of the equation, as though the Kantian conception of reasoning and choice were in fact correct. In matters of education and social policy, how best to factor in feeling and affect is something requiring a great deal of mulling -- and practical wisdom. In any case, my hope is that understanding more about the empirical facts of decision-making, both at the neuronal level and behavioral level, may be useful as we aim for practical wisdom and ponder issues of social policy.

### Acknowledgements:

I am particularly indebted to Hanna Damasio and Antonio Damasio for extended discussions on these and related topics. I wish also

to thank Francis Crick, Paul Churchland, Rodolfo Llinas, David Brink, Deborah Forster, Jordan Hughes, Philip Kitcher, Laura Reider, and the members of the Experimental Philosophy Lab at UCSD.

REFERENCES:

- Aristotle. **The Nichomachean Ethics**. translated by J. A. K. Thompson (1955). Harmondsworth: Penguin Books.
- Bechara, A., A. R. Damasio, H. Damasio, and S. W. Anderson. (1994). "Insensitivity to future consequences following damage to human prefrontal cortex." **Cognition**. 50: 7-15.
- Campbell, C. A. (1957). **On Selfhood and Godhood**. London: George Allen and Unwin Ltd, and New Jersey: Humanities Press Inc. pp. 158-179.
- Cheney, D. L. and R. M. Seyfarth (1990). **How Monkeys See the World**. Chicago: University of Chicago Press.
- Churchland, Paul M. (1995) **The Engine of Reason, The Seat of the Soul**. Cambridge, MA: MIT Press.
- Clutton-Brock, T. H. and G. A. Parker (1995). Punishment in animal societies. **Nature**. 373: 209-216.
- Crick, Francis. (1994) **The Astonishing Hypothesis**. New York: Scribners.
- Damasio, Antonio (1994). **Descartes' Error**. New York: Grosset/Putnam.
- Damasio, A. R., D. Tranel, and H. Damasio (1991). Somatic markers and the guidance of behavior. In: **Frontal Lobe Function and Dysfunction**. Edited by H. Levin, H. Eisenberg, and A. Benton. New York: Oxford University Press.
- De Sousa, Ronald. (1990) **The Rationality of Emotion**. Cambridge, MA: MIT Press.
- Donagan, Alan (1977). **The Theory of Morality**. Chicago: Chicago University Press.
- Flanagan, Owen (1991). **Varieties of Moral Personality: Ethics and Psychological Realism**. Cambridge: Harvard University Press.
- Gewirth, Alan (1978). **Reason And Morality**. Chicago: Chicago University Press.
- Hobart, R. E. (1934). "Free will as involving determinism and inconceivable without it." **Mind**, 43, pp. 1-27.
- Hume, David. (1739). **A Treatise of Human Nature**. Edited by L. A. Selby- Bigge as Hume's Treatise. Oxford: Oxford University Press.
- Johnson, Mark (1993). **Moral Imagination**. Chicago: Chicago University Press.
- Kagan, Jerome. (1994). Galen's Prophecy: **Temperament in Human Nature**. New York: Basic Books
- Kant, Immanuel (1797). "Fragments of a moral catechism". In: **The Metaphysical Principles of Virtue**, translated by James Ellington (1964). New York: Bobbs-Merrill. pp. 148-53.
- Kenny, A. J. P. (1989). **The Metaphysics of Mind**. Oxford: Clarendon Press. Kluver, H. and P. C. Bucy (1937). "Psychic blindness' and other symptoms following bilateral temporal lobectomy in rhesus monkeys. **American Journal of Physiology**, 119: 352-352.
- Kluver, H. and P. C. Bucy (1938). "An analysis of certain effects of bilateral temporal lobectomy in the rhesus monkey, with special reference to 'psychic blindness'". **Journal of Psychology** 5: 33-54.
- Libet, Benjamin (1985). "Unconscious cerebral initiative and the role of conscious will in voluntary action." **The Behavioral and Brain Sciences**. 8:529-566.
- MacLean P. D. (1949) "Psychosomatic disease and the 'visceral brain'. Recent developments on the Papez theory of emotion". **Psychosomatic Medicine** 11: 338-353.
- MacLean, P. D. (1952) Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain). **Electrophysiological and Clinical Neurophysiology**. 4: 407-418.
- Nagel, Thomas (1970). **The Possibility of Altruism**. Princeton University Press.
- Nicholls, J. G., A. R. Martin, and B. G. Wallace (1992). **From Neuron to Brain**. Sunderland MA: Sinauer Associates.
- Papez, J. W. (1937). "A proposed mechanism of emotion." **Archives of Neurology and Psychiatry**. 38: 725-744.
- Piercy, Marge (1982) **Braided Lives**. New York: Knopf.
- Rawls, John. (1971) **A Theory of Justice**. Cambridge, MA: Harvard University Press.
- Saver, J. L. and A. R. Damasio (1991). "Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. **Neuropsychologia**, 29: 1241-1249.
- Schlick, Moritz (1939). "When is a man responsible?". In: **Problems of Ethics**, translated by David Rynin. New York: Prentice-Hall, pp. 143- 156.
- Spinoza, Baruch (1677) **Ethics**. Republished in *The Collected Works of Spinoza*, ed. by E. Curley. Princeton, NJ: Princeton University Press.
- Taylor, Richard. (1974). **Metaphysics**. Englewood Cliffs, New Jersey: Prentice-Hall Inc.
- de Waal, F. B. M. (1982). **Chimpanzee Politics**. London: Allen and Unwin.
- de Waal, F. B. M. (1989). "Dominance 'style' and primate social organization." In: **Comparative Socioecology: The Behavioral Ecology of Humans and Other Mammals**. Ed. by V. Standen and R. Foley. Oxford: Blackwells.
- Wittrup, Eleanor. (1994). **A Mind With a Heart of its Own**. Ph.D. Dissertation for UCSD. (unpublished)

Patricia Smith Churchland  
Philosophy Department 0119  
UCSD  
La Jolla CA 92093