

DRAFT**Patricia S. Churchland****Philosophy Department****UCSD****Chapter 5 of Brain-Wise: Studies in Neurophilosophy (in press) MIT Press.****Free Will****5.1 Introduction¹**

Much of human social life depends on the expectation that agents have control over their actions and are responsible for their choices. In daily life it is commonly assumed 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 member of a social species, we recognize co-operation, loyalty, honesty, and helping as prominent features of the social environment. We react with hostility when group members disappoint certain socially significant expectations. Inflicting disutilities (e.g. shunning, pinching) on the socially erring and rewarding civic virtue help restore the standards.

In other social species too, social unreliability, such as a failure to reciprocate grooming or food-sharing, provoke a reaction likely to cost the erring agent. In social mammals at least, mechanisms for keeping the social order seem to be part of what evolution has bequeathed to our brain circuitry. The stability of the social-expectation

¹ Portions of this chapter are drawn from P. S. Churchland (1996b).

baseline is sufficiently important to survival that individuals are prepared to incur some cost in enforcing those expectations. Anyone with dogs can observe the complex but general phenomenon of maintaining social stability in dog interactions. Mature dogs will teach pups what is unacceptable conduct, and typically dogs test newly encountered dogs, making it clear what territory is theirs and what person they will defend. Just as anubis baboons learn that tasty scorpions are to be found under rocks but cannot just be picked up, so they learn that failure to reciprocate grooming when it is duly expected may incur a slap. As discussed in Chapter 3, much of our behavior is guided by the expectation of specific consequences of events, not only in the physical world, but also in the social world. (Figure 5.1)

If the reward and punishment system is to be effectively engaged in shaping social behavior, the actions for which the agent is rewarded or punished must be under the agent's control. The important question, therefore, is this: What is it, for us or baboons or chimpanzees, to have control over our behavior? Are we ever *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 the places where issues about free will bump up against practical reality and our developing negotiation and understanding of what is fair, what is reasonable, and what is effective.

5. 2 Are we responsible and in control if our choices and actions are caused?

One 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 falls

on me and I hit you, then my hitting you was caused by the falling body; I did not choose to hit you. I am not responsible, therefore, for hitting you. Were you to punish me for hitting you, it would not help me avoid such events in the future. Examples emanating from this prototype have been extended to the broader idea that for *any* choice to be free, it must be absolutely *uncaused*. That is, it is suggested that I make a free choice when, without any prior cause and without any prior constraints, I make a decision that results in an action. Examples allegedly illustrating freely chosen actions are Eisenhower's decision send troops into Little Rock to enforce school desegregation, or my decision to go to the coffee shop for a cappuccino. This *contra-causal* construal of free choice is known as libertarianism.² Is it plausible? That is, are the paradigm cases of free choices actually *uncaused* choices?

As Hume demonstrated in 1739³, the answer is *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. Thus Eisenhower's decision was the outcome of his beliefs about the situation and his desire to ensure that the federal school integration law was not flaunted. His decision did not suddenly spring uncaused into existence without preceding beliefs, thoughts, hopes and worries. I went to get a cappuccino because I usually have one about this time in the afternoon, I wanted to have one, and I knew I had enough money to pay for it, and so on. Save for these causal antecedents, albeit *cognitive* causal antecedents, I would not have gone for coffee. By contrast, suppose that without any antecedent causes, I suddenly enter a saloon, ask for a glass of vodka, and gulp I down. I had no antecedent desire for vodka, no habit of going to a

² See Campbell (1957), Kane (1996)

saloon anytime, let alone in the afternoon, and the behavior would be considered utterly at odds with my cognitive state and temperament. Is *this* the paradigm of free choice? Is *this* prototypically responsible behavior?

Reflecting on these sorts of possibilities, Hume made the deeper and more penetrating observation that an agent's choices are not considered freely made *unless* they are caused by his desires, intentions, and so forth. Randomness, pure chance, utter unpredictability, are not preconditions for attribution of responsible choice. Hume puts the matter 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."⁴

Logic reveals, Hume argued, that responsible choice is actually *inconsistent* with libertarianism (uncaused choice). Someone may choose to climb onto his roof because he does not want the rain to come into his house, he wants to fix the loose shingles that allowed the rain in, 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, though he may not be introspectively aware of them as causes. If, without any determining desires and beliefs, he simply went up onto the roof -- *for no reason* -- his sanity and hence his control are seriously in doubt.

³In his anonymously published, *A Treatise on Human Nature*. See the edition by Selby-Bigge (19@@).

⁴ Hume, p. 411

More generally, a choice undetermined by anything the agent believes, intends, or desires is the kind of thing we consider *out* of the agent's control, and is not the sort of thing for which we hold someone responsible. Furthermore, desires or beliefs that are uncaused (if that is physically possible), rather than caused by other stable features of the person's character and temperament, likewise fail to be the conditions for responsible choice. If a desire suddenly and without no antecedent connection to my other desires or my general character were to spring into my mind -- say, the serious desire to become a seamstress -- I would suspect that someone must be "messing with my mind". The brain presumably has no mechanism for introspectively recognizing a desire to fix the roof as a cause, just as it has no way of detecting in introspection that growth hormone has been released or that blood pressure is at 110/85. Nevertheless, a desire most certainly is a cause.

Neither Hume's argument that choices are internally caused nor his argument showing that libertarianism is absurd have ever been convincingly refuted. Notice, moreover, that his arguments hold regardless of whether the mind is a separate *Cartesian substance*, or is pattern of activity of the physical brain. And they hold regardless of whether the etiologically relevant states are conscious or unconscious.

In fact, however, the brain does indeed appear to be a causal machine. So far, there is no evidence at all that some neuronal events happen without any cause. True enough, neuroscience is still in its early stages, and we cannot absolutely rule out the possibility that evidence will be forthcoming at some later stage. Given the data, however, the odds are against it. Importantly, even were uncaused neuronal events to be discovered, it is a further and substantial matter to show that precisely *those* events

constitute choice. They might, for all we can tell know, have to do with features of growth hormone release or variations in the sleep/wake cycle.

Though all events in the brain may be caused, this does not imply that events are predictable. Causality and unpredictability are entirely compatible. Causation concerns conditions that bring about an event, whereas predictability concerns what we *know* about such conditions. When an event occurs in a complex system, we may know that event is causally governed, even though on any given occasion we may be unable to predict precisely the nature of the event. Moreover, we can often make general predictions even if precise predictions are impossible. Thus I might be able to predict that a dollar bill dropped from the top of Eiffel Tower will fall to the ground, but I will be unable to predict exactly the fluttering pattern and its precise downward trajectory. For that will depend on moment-by-moment changes in air currents, and which will occur much faster than I can take relevant measurements and do the computations, even if I were lucky enough to have very powerful computational equipment. Every movement of the dollar bill is, nonetheless, caused.

Similarly, brain events are probably all caused events, but this does not imply that I can predict with any great precision what you will say if I ask you for directions from UCSD to the Salk Institute. I can predict roughly what you will say, however, if I know that you are familiar with the area, that you are alert, paying attention, are not easily disoriented, and that you tend to be forthcoming when asked for directions. I can also predict with considerable confidence that given the opportunity, a human will go to sleep at night for at least a few hours, that he will want to eat and drink at some time during a twenty-four hour period, that he will not want to sit for very long naked on an

iceberg, and so on. I can predict that a neonate will suckle, a puppy will chew shoes and that most undergraduates will name carrots as the first vegetable that comes into their minds. But these are rough and general, not precise, predictions.

The brain is a dynamical system of enormous complexity. The human brain is calculated to have about 10^{12} neurons, and about 10^{15} synapses. The time scale for neuronal events is in the millisecond range. Assuming the synaptic events and neuronal events are the only causally relevant events, then to a first approximation, this means that the human brain has about 10^{15} parameters that can vary over roughly 1-100 milliseconds. (This is a conservative estimate, since there are intraneuronal events, such as gene expression that are also relevant.) These figures mean that it is not physically possible to take all the relevant measurements and perform the relevant computations to grind out a precise prediction in real time. So predictability based on a neuron-by-neuron or synapse-by-synapse basis is even less realistic than predicting the precise path and flutter of the dropped dollar bill. The point of logic, therefore is this: Causality does not entail predictability, and *un*predictability does not entail noncausality. Put another way, causality and unpredictability are entirely consistent.

As we reflect on what would have to be true in order for us to have free choice, we may become unjustifiably impressed with the fact that absolutely precise prediction of an agent's behavior is really impossible. We nurture that hunch that if you cannot predict whether I will choose a green salad or a beet salad, or whether I will choose to say "hi" or "good morning", then my choices are uncaused and therein lies my freedom to choose. The hunch may be the more compelling if it gets support from this tacit

assumption: since 'uncaused' implies 'unpredictable', then 'unpredictable' implies 'uncaused'. As I have shown, however, the implication goes only one way.

'Unpredictable' does *not* imply 'uncaused'. Once the logic of the relation between causality and predictability are clarified, no logical rationale remains for deriving expectations of noncausality from facts of unpredictability.

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. With the advent of quantum mechanics and the respectability of the idea of quantum indeterminacy, the suggestion that somehow or other quantum-level indeterminacy is the basis for a "solution" to the problem of free will remains attractive to some libertarians.⁵ Stripped to essentials, the hypothesis claims that although an agent may have the relevant desires, beliefs etc., he still can make a choice that is truly independent of all antecedent causal conditions. On this view, the agent, not the agent's brain or his desires or his emotions, freely chooses between cappuccino and latte, for example. It is at the moment of deciding that the indeterminacy or the noncausality or the break the in the causal nexus – whatever one wants to call it – occurs. The subsequent choice is therefore absolutely free.

This is meant to be an empirical hypothesis, and as such, it needs to confront neurobiologically informed questions. For example, what exactly, in neural terms, is the "*agent who chooses*"? How does that fit with what we understand about self and self-representational capacities in he brain? Under exactly what conditions do the supposed

⁵ See for example, Kane (1996) and Stapp (1999). For more discussion, see also Walter (2000)

noncaused events occur? Does noncausal choice exist only when I am dithering or agonizing between two equally good – or perhaps equally bad – alternatives? What about when, in full conversation, I use the word "very" rather than the word "extremely"? Does it exist with respect to the *generation* of desires? Why not? There are also questions from quantum physics, such as this: what is the mechanism of amplification of the nondeterministic events? Were quantum effects of the envisioned kind to exist, how could they fail to be swamped by thermal indeterminacy?

These are just the first snowballs in an avalanche of empirically informed questions. Part of their effect is to expose the flagrantly *ad hoc* character of the hypothesis. That is, it is based more in a desire to prop up a wobbling ideology than in factual matters. Rather than fully discussing its merits and flaws now, however, we shall defer a closer analysis of the hypothesis of a quantum-level origin for uncaused choice until further details of the neurobiology of decision-making are on the table. That will allow us to see what bearing the neurobiological data have on the question of causality and choice in the brain, and hence provide a richer context for evaluating the hypothesis of noncausal choice. We return to this hypothesis, and its critics therefore, in Section 5.5 to see how it fares.

Provisionally, therefore, let us adopt the competing hypothesis, namely that Hume is essentially right, and all choices and all behavior *are* caused, one way or another. The absolutely critical point, however, is that not all kinds of causes are inconsistent with free choice; not all kinds of causes are equal before the tribunal of responsibility. Some causes excuse us from culpability; others make us culpable

because they are part of the story of voluntary action. The important question is what are the relevant differences among causes of behavior that makes some kinds play a role in free choice, and others play a role in forced choice. That is, are there systematic *brain-based* differences between voluntary and involuntary actions that will support the notion of agent responsibility? This is the crucial question, because we do hold people responsible for what we take to be *their* actions. When those actions are intentionally harmful to others, punishment, varying from social disapproval to execution, may be visited upon the agent. When, if ever, is it fair to hold an agent responsible? When, if ever, is punishment justified?

Many possibilities have been explored to explain how the notions of control and responsibility can make sense in the context of causation. These fall under the general rubric of “Compatibilism”, which means that our work-a-day notion of responsibility is, at bottom, *compatible* with the probable truth that the mind-brain is a causal machine. First we shall consider some obvious but unsuccessful attempts at squaring responsibility and causation, and then we shall raise the possibility that increased understanding of the brain will aid in piecing together a plausible account.

5.3 Caused Choice and Free Choice: Some Traditional Hypotheses

5.3.1 Voluntary causes are *internal* causes

Can we rely on the following rule: you are responsible if the causes are internal, otherwise not? No, for several reasons. A patient with Huntington's disease makes nonpurposeful, jerky movements as a result of internal causes. But we do not hold the

Huntington's patient responsible for his movements, since they are the outcome of a disease that causes destruction in the striatum. He has no control over his movements, they are not voluntary, and they are not consistent with his actual desires and intentions, which he cannot execute. A sleepwalker may unplug the phone or kick the dog. Here too the causes are internal, but the sleepwalker is not straightforwardly responsible. In a rather attenuated sense, the sleepwalker may *intend* his movements, though he is apparently unaware of his intentions.

5.3.2 Voluntary causes are *internal*, they involve the agent's intentions, and the agent must be aware of his intention

This strategy also fails. A patient with obsessive-compulsive disorder may have an overwhelming urge to wash his hands. He wants and intends to wash his hands, and he is fully aware of his desire and his intention. He knows that the desire is his desire; he knows that it is he who is washing his hands. Nevertheless, in patients with obsessive-compulsive disorder (OCD), obsessive behavior such as hand-washing or footstep-counting is considered to be out of the agent's control. They often indicate that they wish to be rid of hand-washing or footstep counting behavior, but cannot stop. Pharmacological interventions, such as Prozac, may enable the subject to have what we would all regard as normal free choice about whether or not to wash his hands.

5.3.4 Voluntary causes *feel* different from the inside

Another strategy is to base the distinction between voluntary causes and involuntary causes 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 and help. Is introspection a reliable guide to responsibility? Can introspection -- attentive, careful, knowledgeable 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, the OCD patients just mentioned, and patients with Tourette's syndrome are obvious examples that muddy the waters. In a patient with claustrophobia, the desire not to go into a cave feels as much *his* as his desire not to go rafting without a life jacket. He can even give reasons for both -- it could be unsafe, avoidable injuries could happen, etc. His desire not to go into a cave may be very strong, but so may his desire to eat when hungry or sleep with his wife. So mere *strength* of desire will not suffice to distinguish actions for which the agent has undiminished responsibility and those for which he is not fully responsible.

The various kind of addictions present a further range of difficulties. A smoker feels that the desire for a cigarette is indeed *his*. His reaching for a cigarette may feel every bit as free as reaching to turn on the television or scratching his nose. He might wish it were not his, but so far as the *feeling* itself is concerned, it is as much his as his desire to quit smoking. 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 all of that interest, inclination, and alteration of behavior *feels* -- from the inside at any rate -- entirely free.

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 by others 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 the fact, however, that what is in fashion has a huge effect on what we find beautiful, and this affects not only our choices of clothes, but also such things as aesthetic judgment regarding plumpness or slenderness of the female body. Baseball hats worn backwards have been in fashion for about ten years and are considered to look good, but from another perspective, most people look less attractive if wearing a baseball cap backwards.

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 places ten pairs of identical panty hose, and asked shoppers to select a pair, and then briefly explain their choice. 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. None of them considered this to be a factor, none of them referred to it as a basis for choice, yet it clearly was so. The ten pairs of panty hose were, after all, identical to one another. Other examples of priming, subliminal perception, and emotional manipulation, also suggest that we will not get very far with appeals to introspection to solve our problem about which behavior is in our control and which is not.

5.3.5 Could have done otherwise

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 is, in some sense,

the agent had the power to do something else.⁶ Certainly this idea does comport with conventional expectations about voluntary behavior, and insofar, it is appealing. Lyndon Johnson, historians say, could have done otherwise regarding Vietnam. He could have decided to stop the war in Vietnam in 1965 when he correctly judged it to be unwinnable. I could have decided not to get coffee, and perhaps to have water instead. Nobody *forced* me or coerced me; the desire for coffee was mine. So far so good. 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 including such unpredicted and undirected outbursts as, "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, the results obviously would have been different. Nevertheless, we hold the parliamentarian responsible, but not the Touretter. So the proposed criterion seems not so much wrong, as unhelpful in revealing the nature of the difference between the causes of voluntary behavior and the causes of *nonvoluntary* behavior.

The further problem lurking here is circularity. Testing for whether an agent could have done otherwise seems to be exactly the same as testing whether the behavior was voluntary. Hence specifying what counts as voluntary behavior by referring the

⁶ (See R. Taylor (19@@); Kenny (1989))

possibility that the agent might have done otherwise just goes round in a small circle. It does not seem to get us anywhere.

5.4 Towards a Neurobiology of Decision-Making and Free Choice

5.4.1 Prototypes and Responsibility

In our legal as well as daily practice, we accept certain prototypical conditions as excusing a person from responsibility, but assume him responsible unless a definite exculpatory condition obtains. In other words, responsibility is the default condition; excuse from and mitigation of responsibility has to be positively established. The set of conditions regarded as exculpatory can be modified as we learn more about behavior and its etiology. A different but related issue concerns what to do with someone who harms others but has diminished responsibility.

Aristotle (384-322 BC) in his great work, The Nicomachean Ethics, was the first to articulate that one is responsible unless there are exculpating reasons. And wise principle is, so wise that the core of his approach is still reflected in much of human practice, including current legal practice. In his systematic and profoundly sensible way, Aristotle pointed out that for an agent to be held responsible, it is necessary that the cause of the agent's behavior be internal to the agent. 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, 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. A captured soldier is supposed to resist giving information to the enemy. As Aristotle illustrated in his own discussion of such complexities, we seem to deal with these cases by judging their similarity to uncontroversial and well-worn prototypes, which is perhaps why precedent law is so useful.⁷

Increasingly, it seems unlikely that there is a sharp distinction -- brain-based or otherwise -- 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. This implies not that there is no distinction, but only that whatever the distinction, it is not sharp. That is, it is not like the distinction between having a valid California divers' license and *not* having a valid California drivers' license. It is rather more like categories with a prototype structure; e.g. "being a good sled dog", "being a river navigable by canoe", "being a fertile valley". These sorts of categories are useful even though we cannot specify necessary and sufficient conditions for membership in such categories, but teach them by citing prototypical instances, along with contrasting prototypical *non*instances.

Once we consider being in control in this light, we instantly recognize the degrees and nuances typical of freedom of choice. An agent's decision to change television channels may be more unconstrained than his decision to pay for his child's college tuition, which may be more unconstrained than his decision to marry his wife, which may be more unconstrained 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-

⁷ See more extended explanations in P. M. Churchland (1995)

control in some circumstances than in others. Prolonged sleep deprivation makes it extremely hard to stay awake, even when the need to do so is great. Hormonal changes, for example in puberty, make certain behavior patterns highly likely, and in general, the neurochemical milieu has a powerful effect of the strength of desires, urges, drives and feelings.

These considerations motivate thinking of control as coming in degrees, and hence as falling along a spectrum of possibilities. Towards opposite ends of the self-control spectrum are prototypical cases that contrast sufficiently in behavioral and internal features to provide a foundation for a basic, if somewhat rough-hewn, fuzzy-bordered distinction between being in control and not, between freely choosing and not, between being responsible and not. In fact, as we consider various points on the spectrum, it seems likely that there are in fact *many* parameters relevant to being in control. Consequently, we should upgrade the simple one-dimensional notion of a *spectrum* to a multi-dimensional notion of a *parameter space*, where the dimensions of the parameter space reflect the primary determinants of in-control behavior.

In our current state of knowledge, we do not know how to specify all the parameters or how to weight their significance. And the relations among the parameters are not likely to be linear. We can nevertheless make a start. We do know now that activity patterns in certain brain structures, including the anterior cingulate cortex, hypothalamus, insula, and ventromedial frontal cortex are important. For example, large bilateral lesions to anterior cingulate abolish voluntary movement, though the patient remains aware of his surroundings.⁸ One fortunate patient recovered some voluntary

⁸ This syndrome is also known as *akinetie mutism*. For a review paper, see Vogt, Finch and Olson (1992)

function after a period of ininitiiion. She also had good memories of her symptomatic epidose, during which, she explained "nothing mattered" and that she said nothing because she "had nothing to say".⁹ Smaller lesions to the anterior cingulate are associated with severe depression and anxiety.¹⁰ (Figure 5.2)

If a lesion occurs in the middle area of the cingulate cortex, patients may show loss of voluntary control over a hand. In *alien hand syndrome*, as this deficit is called, the hand behaves as though it has a will of its own. To the consternation of the patient, the hand may grab cookies, or behave in socially inappropriate ways. One patient discovered he could regain some control over his misbehaving alien hand if he yelled at it, "stop that!"

Imaging data implicate the anterior cingulate gyrus in the exercise of self-control over sexual arousal. In an fMRI study, male subjects were first exposed to erotic pictures, and then were asked to inhibit their feelings of sexual arousal. Comparisons between the two conditions show that when subjects are responding normally to erotic pictures, limbic areas show increased activation. When subjects engage in inhibition of sexual arousal, this activation disappears, and the right anterior cingulate gyrus and the superior frontal gyrus become more highly activated. ¹¹

The anterior cingulate again emerges as a player in autism. One undisputed finding is that autistics have deficits in analyzing affective signals. Because limbic structures play a central role in affect, a leading hypothesis claims that autism results primarily from defective affective evaluation, owing to from structural abnormalities in

⁹ Damasio and Van Hoesen (1983)

¹⁰ Balantine, et al. (1987)

¹¹ For this study, see Beauregard, Lévesque, and Bourgouin (2001)

limbic system.¹² This hypothesis has been tested by comparing the microstructure of normal and autistic brains. Using whole-brain serial sections, researchers examined the brains of nine deceased autistic subjects. The only cortical structure to show abnormalities was the anterior cingulate gyrus, where the cells were smaller and the packing density greater. There were similar abnormalities in limbic subcortical structures including the hypothalamus, amygdala and mammillary bodies. Abnormalities in the cerebellum were also seen.¹³

Additionally, it is known that levels of neuromodulators, such as serotonin and dopamine, and of neurotransmitters such as norepinephrine and acetylcholine, as well as various hormones such as estrogen and testosterone, are highly pertinent parameters in the well-tuned decision-making neural organization. For example, obsessive-compulsive pathologies and depressive pathologies involving loss of motivation can be greatly modified by up-regulated serotonin levels. (Figure 5.3) It is also known that subjects with Klinefelter's syndrome (that is, with XXY chromosomes) have poor long-term judgment and impulse control, even when they are cognitively capable. Yet the judgmental capacities of Klinefelter's subjects improve markedly when they are given constant administration of testosterone through a skin patch. Tourette's syndrome is much more controlled when patients are given serotonin agonists; the subjects just do not feel the same desire to engage in their customary *ticcing* behavior. Since the anterior cingulate has been implicated in voluntary behavior, it is noteworthy that both the dopamine projections and the norepinephrine projections can influence

¹² For a fuller discussion, see Hobson (1993)

¹³ See Bauman and Kemper (1995)

activity in the anterior cingulate, and thus have an influence on executive and attentional functions.¹⁴ (See Figure 5.4 Robbins and Everitt)

Appetite is a particularly promising parameter to consider in discovering the brain-based differences between being or not being *in control*. Gluttony allegedly is one of the seven deadly sins; overeating, we are repeatedly reminded, can be controlled by sheer will power. The discovery of the role of the protein leptin in eating, and particular in over-eating, has provoked reconsideration of just how much freedom of choice to push back from the table the very obese actually have, and whether leptin-related interventions will give them greater control.¹⁵

Leptin is a hormone released by fat cells. It acts on neurons in the hypothalamus that regulate feelings of hunger and satisfaction. Experiments on normal mice show that when the mouse has had an adequate meal, the leptin levels *increase*, and the mice leave the food for other pleasures. Some mice are different. They are obese, and they continue to eat even when their leptin levels rise. Genetic analysis shows that the receptor to which leptin binds can have a variety of mutations, and that the specific mutation predicts how overweight the animal is. For example, if the mouse has the *tu* mutation it is somewhat tubby, relative to normals, and has *twice the leptin levels of normals*; if it has the *db* mutation, it is truly obese, and has ten times the leptin levels. There is something *very* different about the appetite regulation of the mutant animals. (Figure 5.5 shows the reward pathways.)

¹⁴ For a review paper on the ascending projection systems, see Robbins and Everitt (1995)

¹⁵ This was first pointed out to me by Carmen Carillo in a paper for my class, and was subsequently discussed in an editorial in *Nature neuroscience*, Fat and free will (2000), 3:1057. [what is the full ref for the refs section?](#)

If one is born with the db mutation of the leptin-receptor gene, and if, in consequence, one feels as ravenous at the end of dinner as at the beginning, it seems inevitable that one will overeat. More precisely, it seems reasonable to assume that such a person will have less control over his eating behavior than a person with the standard version of the leptin-receptor. He may have perfectly normal self-control when it comes to other matters, such as sex, alcohol or gambling, but for food, his situation is markedly different because his leptin-receptors in the hypothalamus are markedly different. The suggestion, therefore, is that the leptin-receptor and its possible variations constitutes yet another component in the complex neurobiological profile of “in control” subjects, at least where food is concerned.

Many neural details remain to be uncovered, needless to say, but identifying the major neurochemical players is a profoundly important beginning. Beginnings such as these inspire the vision that neuroscience might ultimately be able specify a range of optimal values for the relevant parameters. When values fall within the optimal range, the agent's behavior is in his control. When values are suboptimal, the agent will be unable to control his behavior. In between, there may be gray areas where the agent is neither fully in control nor fully out of control.¹⁶ (Figure 5.6)

Research from basic neuroscience as well as from lesions studies and scan studies will be needed to transform this speculative parameter space into a substantial, detailed, testable account of the features that are typical of “in control” subjects. These properties may be quite abstract, for “in control” individuals may have different

¹⁶ See also Walter (2000)

temperaments and different cognitive strategies.¹⁷ 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 well or poorly certain such tasks as walking. What I am proposing here is that more abstract skills, behaviorally characterized, 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. (Figures 5.7 and 5.8)

In the next sections, we shall consider in more detail some of the evidence that speaks in favor of this general approach.

5.4.2 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 and minimally emotional. To achieve rationality and self-control, one must maximally the suppress emotions, feelings, and inclinations. In a metaphor sympathetic

¹⁷ See Kagan, *Galen's Prophecy* (1994).

to this idea, Plato characterizes reason as a charioteer who is pulled along by the appetites and emotions, and who must beat them to avoid running amok.

Immanuel Kant is the philosopher best known for insisting on an opposition between reasons and emotions, and favoring the supremacy of reason. In his moral philosophy, Kant saw human agents as attaining virtue only as they succeed in downplaying feeling and inclination. He says: "The rule and direction for knowing how you go about [making a decision],¹⁸ without becoming unworthy of it, lies entirely in your reason. This amounts to saying that you don't learn this rule of conduct by experience or from other people's instruction; your own reason teaches and even tells you what to do." The perfect moral agent, on Kant's view, is perfectly rational and entirely without emotion and feeling.¹⁹ (Ronald de Sousa calls such an agent a "Kantian monster"²⁰)

The kinds of cases that inspire Kant's veneration of reason and his suspicion of the passions are the familiar "heart-over-head" blunders. In such cases, the impassioned do-gooder makes things worse, or one neglects long-term consequences while satisfying an immediate need. The fool does not look before he leaps. Othello, so overcome by jealousy that he failed to realize that he was being duped, kills Desdemona. In the grip of an overwhelming bitterness, Medea kills her children and herself. The moral failings of great tragedy are typically character flaws involving great emotions engulfing weak reason.

¹⁸Kant actually says, "The rule and direction for knowing how you go about *sharing in happiness...*" (my italics) because the matter arises in the context of a teacher-student dialogue about a particular case, namely how to help others and whether to give them what they want. Pretty clearly Kant intends the point to be general, and hence my more general interpolation. **Which Kant Ref? Is it in refs section?**

¹⁹ Or as Marge Piercy remarks in *Braided Lives* (1982) "...treats his emotions like mice that infest our basement or rats in the garage, as vermin to be crushed in traps and poisoned with bait."

Understanding the consequences of a plan, both its long term 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. 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." (1739) 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." As Hume understands it, reason is responsible for delineating the various *consequences* of a plan, and thus reason and imagination work together to anticipate pitfalls and payoffs. But feelings, informed by experience, are generated by the mind-brain in response to anticipations, and incline an agent for 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 severally portrayed as prototypically hot-tempered, or coldly reasonable, or moderate in all things. The pointy-eared semi-alien, Mr. Spock, lacks emotion. In trying circumstances, his head is cool and his approach is calm. He faces catastrophe and narrow escape with comparable equanimity. He is puzzled by the humans' propensity to anger, fear, love and sorrow, and correspondingly fails to predict

²⁰ See de Sousa (1990), p. 14

the role of emotions in human affairs. 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, ignore future costs, or 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 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 Aristotle's ideal of someone who is wise in practical matters.

5.4.3 A Disconnection Effect: EVR

Neuropsychological studies reveal a lot about 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 impractical and disadvantageous in the long run. Thus SM, whose amygdala has been destroyed, has no feelings of fear. (See again Figure 5.1) In complex circumstances, with no access to gut feelings of unease and fear, she is as likely as not to make a decision that normal people could easily foresee to be contrary to her interests. In a rather more complex way, the point is dramatically illustrated by the remarkable man, EVR who first came to the Damasio's lab at the University of Iowa College of Medicine more than a decade ago.²¹

²¹ Damasio (1994)

A brain tumor in the ventromedial region of EVR's frontal lobes had been surgically removed, 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 troubling picture began to emerge. Once a steady, resourceful and efficient accountant, now EVR 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, EVR's problems seemed to his physician more likely to be psychiatric than neurological, and hence best treatable with psychoanalysis. As we now know, the psychiatric diagnosis turned out to be quite wrong.

After studying EVR for some time, the Damasio and their colleagues conjectured that his lapses in practical judgment had something to do with a disconnection between emotions and judgment. They repeatedly observed that although EVR could state 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), 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).²² When they tested whether EVR's emotional responses were in the normal range, they found intriguing abnormalities. For example, when shown horrifying or disgusting or erotic pictures, his galvanic skin

²² Saver and Damasio (1991)

response (GSR)²³ was flat. (Normals, in contrast, show a huge response while viewing such pictures.) Curiously, if asked to say what he saw in the pictures, EVR's emotional responses became somewhat more normal.

During the following years, new and more revealing tests were devised to probe more precisely the relation between reasoning logically on the one hand, and *acting* in accordance with reason on the other. Antoine Bechara, working with the Damasio, developed a particularly revealing test. In this test, generally known as the Iowa Gambling Task, a subject is presented with four decks of cards and told only that his goal is to make as much profit as possible from an initial loan of money. Money can be made and lost as a function of turning over cards, one at a time, from any of the four decks. Subjects are not told how many cards can be played before the game ends (a series of 100) or what the payoffs are from any deck. One has to discover the winning strategy by trial and error. After a card is turned over, the subject is either rewarded with an amount of money, or on some cards, he 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 make a profit 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. (Figure 5.9)

After about 15-20 trials, normal controls typically come to stick mainly with low-paying/low-penalty decks (C and D) and duly make a tidy profit in the long run. In

²³The GSR measures change in conductivity of the skin as a function of increased sweat on the skin,

contrast, subjects with ventromedial frontal damage tend to end the game with a loss. They generally work the high-paying decks, despite the profit-eating penalty cards in those decks. Subjects with brain damage to regions other than ventromedial behave like controls. Yet the ventromedial patients had normal IQ's.

As Bechara et al. note, even after repeated testing on the gambling task, as long as a month or as short as twenty-four hours after, EVR continued heavily to play the losing decks. When queried at the trial end, inevitably he correctly *reports* that A and B are losing decks, and rues his strategy. To put it rather paradoxically, *rationally* EVR does indeed know what the best long-run strategy is, but in exercising choice in actual situations, he goes for short-run gain, incurring long-run loss. 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. Because EVR can articulate well enough the future consequences of alternative actions, the problem cannot be lack of understanding of what might happen. That his "pure reasoning", displayed *verbally*, is at odds with his "practical decision-making" displayed in *behavior*, suggests that the crux of the problem lies with EVR's lack of emotional responsivity to complex plans.²⁴

Additional results came from a deeper analysis of skin conductance data taken by a galvanometer placed on the arm of each subject during the gambling task.²⁵ 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 when they reached for the

which is an effect produced by the sympathetic system of the nervous system.

²⁴ Bechara, et al. (1994); Damasio (1994); Adolphs and Damasio (@@@@)

²⁵ Damasio et al. (in press)

"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 when starting to reach 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. They remain free of any affective guidance. What is so striking, is that for control subjects, choice was 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 likely.²⁶

The significance of nonconscious biasing by emotion 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 by multiplying the probability of each outcome by the value (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 artificial, at least for the ongoing daily activity of actual humans. Probably it is true for at best a small range of highly manageable problems, but even then, we may wonder whether the set of options is delimited after "feeling" brings to awareness mainly the "feels-reasonable" alternatives. Other *possible* alternatives are just never entertained. At any rate, the economists' model is unlikely to come even close to giving

²⁶Benjamin Libet came to a similar conclusion using a very different experimental paradigm. (Libet, 1985)

the whole story of rational choice, though it may be helpful once the set of reasonable alternatives is laid out.

On many occasions, one's brain seems to have things pretty well sorted out before conscious deliberation even begins. For example, in the grocery store, I rarely bother to consider Delicious apples since they are usually punky; I am not fond of eggplant, so I never pause over the eggplant bin. I never consider drinking a can of paint; I never ponder whether to make a fur bathing suit or porridge skis. I never consider beginning logic class with a demonstration of how to milk a cow. And so forth. All these are descriptions of options my brain *could* entertain, but does not.

One lesson taught us by EVR and others with similar lesions (ventromedial frontal) is that whatever rationality in decision-making *actually is*, independence from emotions is not its essence. 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 little about changes the viscera, and hence he is missing important emotional clues that something is foolish or unwise or problematic. His frontal lobes, needed for a complex decision, have no access to information about the emotional valence of a complex situation or plan or idea. Consequently some of EVR's behavior turns out to be foolish and unreasonable.²⁷ The point is not that patients like EVR 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 because the ventromedial frontal region needed for integration of body-state representation and fancy "scenario-spinning" is

disconnected from the "gut feelings". In particular, it probably entails that the capacity to remember relevantly similar occasions with a recollection imbued with evaluative significance, is impaired. Normally, neurons in ventromedial frontal cortex would project to and from areas such as the anterior cingulate cortex, amygdala and hypothalamus that contain neurons signaling body-state values. In patients with destruction of ventromedial cortex, the pathways are disrupted.

An even more worrisome behavioral profile is seen when the prefrontal lesions occur early in development. Anderson and colleagues²⁸ reported on two adults patients whose prefrontal lesions occurred before the age of 16 months. Both scored normally on various intelligence tests, but both were severely impaired in their social behavior. In addition, they also showed defective social and moral reasoning, suggesting that the capacity to acquire moral understanding was itself diminished by the lesion. Whereas EVR and other late-onset lesion patients might do things that are socially inappropriate or foolish, they understand and abide by moral rules. (Figure 5.10)

5.4.4 Agents and Self-Representational Capacities

If the various emotions play an on-going and indispensable role in formulating plans, both long and short-term, how does this fit into the framework for agency, self-representation and consciousness developed in chapters 3 and 4? The answer is best laid out by referring once more to the Grush emulator. As discussed earlier, to a first approximation, the motivation for actions is anchored in the fundamental drives for food, sex, and survival. As plans develop, the imagination generates representations of plan-

²⁷ Is EVR merely showing frontal perseveration? No, because he does score normally on the Wisconsin card-sorting task, in contrast to perseverative patients. For a much more full account, see Damasio (1994). See also Raine, et al. (1998), and Raine, Buchsbaum, and LaCasse (1997)

²⁸ Anderson, Bechara, H. Damasio, Tranel and A. R. Damasio (1999).

sequelae. To these internally driven scenarios, as well as to perceptually-driven representations, emotional responses are generated, via mediation of the brainstem structures, the amygdala and hypothalamus.²⁹ The central function of the emulator is to predict and evaluate consequences of proposed actions. As we saw, the emulator can be employed on-line in making immediate decisions, and off-line for high-level decisions involving longer time scales. The various emotions have a central role in evaluating options and their consequences as threatening, rewarding, dangerous, risky, painful, satisfying, and so forth. If these affective states also represent the difference between ‘threatening to *him*’ versus ‘threatening *me*’, then the states are, on Damasio’s hypothesis, conscious feelings. In the context of acquired cognitive-cum-emotional understanding about the world, neuronal activity in these pathways calls forth certain memories, it directs attention to certain perceptual and imaginative functions, and it imbues certain perceptions with practical significance.

The neural evaluation and assessment of options probably resembles less the clean, step-by-step execution of an algorithm than it does the rough-and-tumble jostling among puppies for access to the food supply. That is, the process whereby neural networks settle into the “next decision” probably involves a kind of competition, and the winning option moves ahead for assignment of detailed movements. To put it crudely in the familiar framework of folk psychology, a desire for immediate gratification can be outweighed by the fear of missing out on a more valuable good in the longer run; the pain of exercise can be endured for the sake of envisioned improvement in skiing

²⁹ An earlier hypothesis related to this view was suggested by Paul MacLean (1949, 1952). He said "As a working hypothesis, it can be inferred that the limbic system is for the "body viscous", a visceral brain that interprets and gives expression to its incoming information in terms of feeling..."(1952). See also James Papez (1937), Kluver and Bucy (1937, 1938)

performance; long and dreary hours in the lab are sustained by the glimmering possibility of satisfying one's curiosity. On those occasions when a weighty decision involves conscious deliberations, we are sometimes aware of the inner struggles, describing ourselves as having conflicting or ambivalent feelings. Some processes in decision-making take longer to resolve than others, and hence the wisdom in the advice to "sleep on" consequential decisions. Everyone knows that sleeping on a heavy decision tends to help us settle into the 'decision-minimum' we can best live with, though exactly how and why, is not understood. Are these longer processes classically rational? Are they classically emotive? Probably they are not fittingly described by our existing vocabulary. They are the processes of a dynamical system settling into a stable attractor.

Introspection, as we know, is a highly limited and fallible guide to the dynamical aspects of these inner processes, and folk psychology is at best a crude interpretive filter in any case. Though introspection gives us some sense of the neural hurly-burly subserving choice, we have little conscious access of its neural nature. Nevertheless, good models of the interplay and competition among parameters, whatever exactly they are, will probably emerge in time.

According to the conventional wisdom, *cognitive* factors are used to predict consequences, while *emotive* factors are used to evaluate the consequences, and the two are entirely separate functions. From the point of view of the brain, however, the situation is not that simple. The very formulation of certain general goals, such as going to college or starting a business, likely employs an inseparable alloy of cognitive-emotive elements. This is also true of more specific goals, such as taking the dogs to

the beach or finding a babysitter. Scene segmentation and pattern recognition in perception are, save for unusual circumstances, shot through with affect and meaning. In momentous decision-making, such as the decision to find the accused guilty or the decision to opt for doctor-assisted suicide, the competition alluded to is never a one-dimensional struggle between reason and emotion, but rather is a complex interplay between *this* cognitive-emotive consortium and *that* cognitive-emotive consortium. The decision to have a latte rather than a cappuccino is, relatively speaking, a completely trivial decision. Our choice really does not amount to a row of pins. Such trivial choices are not, however, the model for those life decisions which mark us as wise or foolish, as impulsive or measured, as lazy or ambitious. Consequently, in developing adequate models of decision-making, we would do well not to make the latte-cappuccino choice the paradigm for choice generally.

5.5 Learning what's reasonable 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, defer gratification, 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 contouring the terrain of the neuronal state space so that behaviorally appropriate trajectories are "well-grooved" or strongly attractive. 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 proper "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.³⁰

The characterization of a choice or an action as *rational* carries a strongly normative component; insofar, it is not sheerly descriptive, in contrast, for example, to describing the action as performed hurriedly or with a hammer. Claiming that 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 kith and kin; that it properly took into account the consequences of the action, both long term and short term. 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 or reasonable actions, as well as 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.³¹ As Paul Churchland has argued, we also learn ethical concepts such as 'fair' and 'unfair', 'kind' and 'unkind', by being shown prototypical cases, and slowly learning to generalize to novel but relevantly similar situations.³²

Peer and parental feedback fine-tune the pattern recognition networks so that over time they come closely to resemble the standard upheld in the wider community.

³⁰ Damasio (2000)

³¹ see Trends in CogSci ????

Nevertheless, as Socrates was fond of showing, articulating those standards is awesomely difficult, even when a person successfully uses the term 'rational', case by case. Discriminating the reasonable from the unreasonable may be a skill, like discriminating whether the river is now navigable by canoe, or whether and how attacking an enemy's position will succeed. Using prototype knowledge, we can see how Scott's skill in conducting his Antarctic exploration was pitiful, while Amundsen's was superb. Making the term 'rational' precise in a way that fulfills the conditions for an algorithm is almost certainly impossible. Failures to program computers to conform even roughly to common sense, or to understand what is relevant, are an indication of the nonalgorithmic, *skill-based* nature of rationality.

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 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.

Frankly 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 Damasio's evidence from their patient SM who, as a result of

³² P. M. Churchland (1995) From a different perspective, McDowell (19@@) ends up arguing for a similar view. See also Casebeer (2000).

amygdala destruction, has no feelings of fear. Although she can identify which *simple* situations are dangerous, this seems for her to be a purely cognitive, nonaffective judgment. Her recognition is poor, however, when she needs to detect the menace or hostility or pathology in complex social or marketing 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 plausibly poor at predicting what will provoke strong sympathy or dread or embarrassment in humans.

Stories, both time-honored one and 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 Beanstalk*), 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, Aristophanes, 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 silly³³; 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. 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, deciding whether to surrender or press on, etc. -- these are usually complex constraint satisfaction problems.

As we deliberate about a choice, we are guided 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" to indicate that they are a response not to external stimuli, but to internally generated representations and recollections.³⁴ 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.

³³ Recall that Buridan's ass was placed midway between two bales of hay, and could not decide which to approach first, and so died of starvation.

³⁴ Damasio (1994) p. 134 ff.

5.6 Uncaused Choice Considered Again

Much of this chapter has focused on the emerging account of the neurobiology of decision-making. The hypothesis on offer is that there are systematic neurobiological differences between being *in control* and being *out of control*, and that these differences can be characterized in terms of fuzzy-bordered subvolumes of the multi-dimensional parameter space. The in-control subvolume of the space may be relatively large, allowing for the fact that in-control humans have different habits, cognitive styles, emotional tone, and so forth. Similarly, the out-of-control subvolume may be very large, reflecting the fact that dysfunction to the reward system may yield an out-of-control profile that is very different from that of a dysfunctional anterior cingulate cortex which in turn is different from that of a degenerating basal ganglia.

As noted in [Section 5.2@](#), there are spirited defenses of a totally different hypothesis, namely that decisions made by in-control subjects are actually *uncaused* decisions, whereas decisions made by out-of-control subjects are *caused*. The most modern variation defends the idea that quantum indeterminacy is at the root, somehow, of uncaused choice. Though briefly introduced earlier, it is time now to reconsider the idea that real choice requires a break in causality milliseconds prior to the emergence of the brain state that constitutes the choice. (p. 5, [Section 5.2](#)) An empirical hypothesis, it deserves to be weighed and evaluated as an empirical hypothesis and compared to the rather different picture of the brain discussed above.

Hume and his arguments aside, the credibility of the noncausal-choice hypothesis depends on whether it can mesh with what is known so far about neurons and nervous systems. Defenders of the hypothesis want it to be *consistent* with existing well-established neurobiological data, not openly clash with the data. The hypothesis is just that among the many details neuroscience has not yet discovered is this fact: for quantum mechanical reasons, voluntary choice is uncaused. Our task here is to ask whether, given what *is* well-established neurobiologically, this appears to be a plausible hypothesis with promising research prospects. The hypothesis classifies a choice as voluntary if and only if it is not caused. Caused choices, on the hypothesis, are deemed not free. As usual, we can begin by raising questions to which the hypothesis should have some noncontrived answers.

Why and how does a break in causality occur just for those particular brain events that supposedly are paradigm cases of choice?. How does the brain work such that a simple behavior in conformity with good habit – routinely putting on my seat belt, for example – *is* caused, whereas choosing a latte rather than a cappuccino after dithering is *not* caused? What prevents the noncausal events from occurring when a nicotine addict reaches for another cigarette or a child sucks its thumb or a highly trained but off-duty spy surveys his fellow passengers for assassins? If, as is entirely likely, the brain events constituting choice are distributed across many neurons, how is the noncausality (quantum indeterminacy) orchestrated across the relevant population? If the brain events constituting choice are uncaused, then what precisely *are* their relations to background desires, beliefs, habits, emotions and so forth? Philosophical fantasies floated in abstraction from the tough and detailed constraints of the real world

have an in-a-single-bound-Jack-was-free quality. Flippant answers to empirically informed questions are of course always possible: “it just works like that” or “magic!”. Unless the hypothesis can interdigitate with neurobiology and cognitive science to come up with nonfrivolous answers to these questions, however, it will continue to look nakedly *ad hoc*.

Before the hypothesis can be taken seriously, it will have to garner some empirical confirmation and survive empirical tests. If uncaused choice is a quantum-level effect, as may be supposed, the aforementioned questions, as well as those raised in Section 5.2, demand empirical answers: under *exactly what conditions* do the supposed noncaused events occur? Does noncausal choice exist only when I am dithering or agonizing between two equally good – or perhaps equally bad – alternatives? How do quantum-level effects know (so to speak) when to occur and when not? Beyond the business of *decisions*, do quantum-level indeterminacies exist with respect to such processes as the generation of *desires*? Or *beliefs*? Why not? How is it they come into play with only some conscious decisions but not others? Does this break in causality occur at the synapse? If advocates of the noncaused decision-making are serious, they will have to do more than wave the flag of quantum-level indeterminacy and claim that in a single bound the choice is free. They will have to get into the business of empirical confirmation.

6. What Happens to the Concept of Responsibility?

We need now to return to the dominant background question motivating this chapter: if choices and decisions are caused, is anyone ever really responsible for his actions? One very general conclusion is provoked by the foregoing discussion. On the whole, social groups work best when individuals are presumed to be responsible agents. Consequently, 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 we match up assignment of responsibility with being in control, and adopt the default assumption that agents have control over their actions. Barring clear evidence that an agent's behavior was in the out-of-control subvolume of parameter space, then the agent is liable to punishment and praise for his actions. This is of course a highly complex and subtle issue, but the basic idea is that *feeling* the social consequences of one choices is a crucial part of socialization -- of learning to be in the give-and-take of the group. It is part of acquiring the appropriate Aristotlean habits.³⁵ *Feeling* those consequences is necessary for contouring the parameter 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 by watching others engage the world, or by hearing about how others engage the world. Similarly, learning about the social world involves direct or indirect 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 also consistent with compassion, kindness and

³⁵ This view can also be found in the classic essays of Hobart (1934) and Schlick (1939)

understanding. In short, I do not want the simplicity of the general conclusion to mask the tremendous subtleties of child-rearing. Nevertheless, if the only known way for "social decency" circuitry to develop requires that the subject generate the relevant feelings pursuant to social pattern recognition, then the responsibility assumption may be preferable to any version of a thorough-going nonresponsibility assumption.

This leaves it open, of course, that under special circumstances agents should be excused from responsibility or be granted 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 S. Quite as obviously, however, the data do *not* show that no one is ever really responsible, that no one is really deserving of punishment or praise. Nor do they show that when life is hard, one is entitled to avoid responsibility. To most of us, the "Twinkie defense" seems a travesty of justice, but so does ignoring someone's massive lesion in the ventromedial frontal cortex.

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 immense caution. When the target is the nervous system, then caution by another order of magnitude is wanted. Still, not taking action is still doing something, and *acts of omission can be every bit as consequential as acts of commission*.

Second, the movie, *Clockwork Orange*, typically conjured up by the very idea of direct intervention by 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? I do not wish to propose specific guidelines to allow, or disallow, any form of direction intervention. Nevertheless, given what we now understand about the role of emotion in reason, perhaps the time has come to give such guidelines a calm and thorough reconsideration. Approaching these questions with a careful Aristotelian determination to be as wise as possible, may be preferable to giving free rein to unreflective self-righteousness. Ideological fervor, on the right or on the left, can often do greater harm than unhurried common sense.

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) the default presumption that agents are responsible for their actions is empirically necessary to an agent's learning, both emotionally and cognitively, how to evaluate the consequences of certain events and the price of taking risks (R. E. Hobart, Moritz Schlick) . 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 thesis seems evident. Consequently, 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 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 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 informed mulling -- and practical wisdom. In any case, my hope is that understanding more about the empirical facts of decision-making, at both the neuronal level and behavioral level, may be useful as we aim for practical wisdom and ponder improvements in our social policy.

REFERENCES:

Aristotle. 1955. *The Nichomachean Ethics*. Trans. by J. A. K. Thompson.

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. Has the self 'free will'? In *On Selfhood and Godhood*, 158-179.

London: Allen and Unwin, and New Jersey: Humanities Press Inc.

Churchland, P. M. 1995. *The Engine of Reason, The Seat of the Soul*. Cambridge,

Mass: MIT Press.

Damasio, A. R. 1994. *Descartes' Error*. New York: Grossett/Putnam.

Damasio, A. R. 1999. *The Feeling of What Happens*. New York: Harcourt Brace.

Dennett, D. C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*.

Cambridge, Mass.: MIT Press

Le Doux, J. 1996. *The Emotional Brain*. New York: Simon and Schuster.

Walter, H. 2000. *Neurophilosophy of free will: From Libertarian illusions to a concept of natural autonomy*. Cambridge, Mass.: MIT Press.