Our Evolving Concept of 'Voluntariness': A Test Case

Paul Sheldon Davies

Introduction

Section 2.01 of the Model Penal Code states that an agent is criminally guilty only if the relevant action was voluntary, but characterizes 'voluntary' in decidedly negative terms:

> A person is not guilty of an offense unless his liability is based on conduct that includes a voluntary act or the omission to perform an act of which he is physically capable. The following are not voluntary acts within the meaning of this Section: (a) a reflex or convulsion; (b) a bodily movement during unconsciousness or sleep; (c) conduct during hypnosis or resulting from hypnotic suggestion; (d) a bodily movement that otherwise is not a product of the effort or determination of the actor, either conscious or habitual.

This way of characterizing 'voluntariness' is remarkably uninformative. The first three conditions illustrate what is *not* voluntary: bodily movements not "determined" by the agent – reflexive movements, sleepwalking, etc. What, then, are the distinguishing properties of movements that are voluntary? We are given only the slightest of hints in condition (d): movements produced by "effort or determination".

However, such non-specificity is not an automatic indictment. Laws are tools designed to fulfill certain functions and some functions can be executed with relatively blunt instruments. This, I surmise, is true of the above characterization of voluntariness. The

relevant conditions are sparsely specified on the assumption that there is enough shared cultural knowledge concerning the causes of human conduct to fill the gaps. The lack of specificity in the law is tolerable, perhaps preferable, because our shared cultural knowledge enables us to apply the law flexibly in light of the particulars of each case.

If so, then the crucial assumption must be something like this: most adult citizens – those likely to serve as jurors and judges – know *that* we are agents who sometimes "determine" their actions and also know *when*, under *what conditions*, our actions are in fact the results of our "determinations". If this assumption is false, if lawyers, judges, and jurors do not know enough to reliably discern actions genuinely determined by the actor from those determined by other factors, then the law is defective.

The question is whether this crucial assumption is true. Do most adult citizens know that we are agents who sometimes determine their actions? Do most know when, under what conditions, our actions result from such determinations? The question is not whether most citizens believe of themselves that they have such knowledge, but whether they in fact have it. I am going to sketch an argument, the conclusion of which is that such knowledge does not obtain and that this portion of the Model Penal Code is therefore defective. I will then step back and ask how a defender of legal pragmatism might respond to the considerations offered in support of this claim.

A Bit of Conceptual Analysis

I begin with a bit of naïve conceptual analysis. The concept 'reasons for acting' might refer solely to factors that rationalize or justify one's action from the first-person point of view. If you ask why I donated money to charity and I cite my desire to help

others in need, my desire appears to function as my reason. Even if my belief that I acted from a desire to help others is false, still, when I cite that desire as my reason, I am reporting the factors that appear motivationally relevant within my deliberative field. At the same time, it is difficult to see how any alleged reason for acting can qualify as a genuine reason if it was causally irrelevant for the action. If I sincerely but falsely believe that I acted from the desire to helps others, then I may report this false belief in trying to rationalize my action, but the 'reason' cannot be a genuine reason precisely because it is not among the things that actually moved me to act.

Perhaps, in light of this ambiguity, we do well to accept the following two claims:

> Consciously-accessible reasons for acting qualify as *genuine* reasons only if they are among the actual causes of our acting,

and

> When we cite our reasons for acting – when we endeavor to rationalize our actions – we express our *sincere beliefs* concerning what we take to be our *genuine* reasons.

If so, then the act of rationalizing of our actions succeeds only if we report our sincere beliefs concerning the actual causes of our actions. Precisely here is where my skepticism exerts its force, because holding a sincere belief concerning what we take to be the actual causes of our actions surely requires relevant evidence. In particular, sincere beliefs concerning the actual causes of one's action require either that

> (A) the agent have plausible evidence concerning the actual causes,

or, at minimum, that

(B) the agent *not* be faced with evidence that such beliefs are prone to error.

Now, as a matter of fact, most of us probably see ourselves as satisfying condition (A). We probably see ourselves as typically knowing our reasons for acting on the basis of plausible evidence. We may also see ourselves as satisfying condition (B), as not being faced with serious evidence that we do not know our reasons for acting. But, as I will now describe, we do indeed have evidence from current sciences of the mind that the minimal condition in (B) and that, as we will see, is the heart of my skepticism.

Skepticism Concerning Human Self-Knowledge: An Overview

Before describing some of the relevant science, I want to give a four-step sketch of my skepticism. This is just a sketch; a fuller explication and relevant evidence will be given shortly.

(1) There is compelling scientific evidence that we sometimes have confident but nonetheless false beliefs regarding our reasons for acting.

The evidence does not show that we are always wrong; nor does it show that we are wrong most of the time (though that might be true). But it does show that, in many instances, we demonstrably are wrong about the reasons we sincerely give for our actions. In addition:

(2) There is compelling scientific evidence that our reason-giving capacities, even when failing to track the causes of our actions, nevertheless cause us to

believe that we know our reasons for acting and thereby fulfill certain social functions.

The idea is simple. If we do not study ourselves scientifically and thus are unaware of the problem described in (1), then our reason-giving activities proceed apace and thereby contribute to interpersonal trust and social cohesion.

Generalizing from (1) and (2), we may further claim:

(3) From the first-person perspective – from the perspective of the deliberating agent – we cannot reliably discriminate between cases in which our beliefs concerning our reasons are correct and cases in which they are not.

There is no reliable phenomenological difference between cases in which the reasons we give are correct and cases in which they are not. If there were reliable means with which to discriminate, then (3) would be false. But so far as I know, we have no such device.

Finally, the claim in (3) plausibly generalizes to the third-person perspective. This is important. If the acting agent is not justified in claiming to know one's reasons for acting, then surely a third-party observer cannot appeal to the agent's expressed reasons to explain the agent's action. A third-party observer, that is, cannot justifiably claim to know whether the agent's action resulted from "effort" or "determination" by the agent. This is where my skepticism poses the most direct threat to the Model Penal Code:

(4) From the third-person perspective – from the perspective of someone attributing "effort or determination" to another agent – we cannot justifiably

claim to know whether the agent's consciously-accessible intentions are genuine casual factors for the relevant action.

Now, this way of expressing my skepticism requires qualification. Sometimes, under controlled experimental conditions, we can be confident we know someone's reason for acting, and it may turn out that those reasons match the reasons given by the agent. So I think it is possible to attribute genuine reasons from the third-person perspective, but only when we have adequate empirical grounds for the attribution.[1] At present, however, the majority of everyday cases are not like that. In most actual cases, even minimal experimental controls are absent. That is why I endorse the skepticism in (4) and the further claim that, in light of this skepticism, the notion of 'voluntariness' in the Model Penal Code is indeed defective.

(1)-(4) also reveal a conflict in the very constitution of our psychology.[2] Some parts of our psychology, those that dispose us to give and ask for reasons, can fulfill their functions whether or not the reasons given are true. And yet, other parts of our psychology, those that dispose us toward scientific inquiry, fulfill their functions by enabling us to discover how things actually work. Conflicts occur when we as scientists discover that we as social animals sometimes endeavor to "justify" our actions by appeal to "reasons" that are false. On the assumption that "reasons" known to be false lose whatever justificatory power they might have had, such conflicts are indeed vexatious and perhaps worse. I will return to such conflicts in my discussion of legal pragmatism.

[1] The same may be true of knowing one's own reasons for acting, though the relevant sorts of experimental conditions may be trickier to accomplish.
[2] I discuss such constitutional conflicts of the human mind for fully in Davies (2009).

Well, the most I can achieve in a short presentation is to indicate the kinds of considerations that support the above steps, especially the skepticism in (4) and the conflicts that arise from the constitution of our minds. I will do this by focusing on considerations that support steps (1) and (2).

In Defense of (1): A Model

Many years ago Michael Gazzaniga performed experiments on patients whose right and left hemispheres had been surgically severed. In one experiment, the left hemisphere was shown a picture of a chicken claw and the right a picture of a snow scene. After that, both hemispheres were presented with an array of objects and the patient was asked to choose objects relevant to what he had seen a few moments earlier. Using his right hand (controlled by his left hemisphere), the patient pointed to a chicken. Using his left hand (right hemisphere), he pointed to a snow shovel. Then the experimenter asked the patient why he chose those particular objects; the patient was asked to give reasons for his action. Now, bear in the mind that our reason-giving capacities are located mainly in our left hemisphere, which means that information concerning the snow scene, contained only in the right hemisphere, was inaccessible to the patient's reason-giving capacities. The left hemisphere, in fact, had no idea why the left hand had pointed to a snow shovel. Yet when the experimenter asked the patient why he chose those two objects, the patient did *not* say "I do not know." Instead, it confabulated a reason. To be exact, the patient said, "Oh that's simple. The chicken goes with the chicken claw, and you need a shovel to clean out the chicken shed."

So one lesson we learn from split-brain experiments is the following:

> The Information-Deprivation-Due-to-Neural-Damage Model: Whenever our left-hemispheric interpreter operates in the absence of causally relevant information that isolated outside the left hemisphere, it invents a "reason" for the agent's action based on information it can access.[3]

It is easy to generalize, however, from split-brain persons to *all* persons, including persons not suffering neural damage, because a great deal of psychological processing occurs outside conscious awareness, or outside whatever form of awareness is required for giving reasons. It is difficult to overstate the importance of this point, for it leads to an analogue of the above model:

> The Information-Deprivation-Due-to-the-Structure-of-Our-Psychology Model: When our conscious, reason-giving capacities operate in the absence of causally relevant, non-conscious information, they invent a "reason" using whatever information happens to be accessible.

When this happens, the "reasons" we give may "rationalize" or render intelligible our actions, at least from the first-person point of view. But because such "reasons" nevertheless fail to represent the actual causes of our actions, they fail to qualify as genuine reasons.

Well, I want to try to convince you that this model – I will refer to it as the Social Psych Model, since the most vivid experiments to date have been done by social psychologists – is plausibly true of all persons. Doing so is my main argument for thesis (1).

---

[3] Gazzaniga 2000 provides an overview of three decades of research on split-brain patients.

In Defense of (1): Social Psych Experiments

Hundreds of priming experiments demonstrate that a wide range of behaviors – motor, memory, intelligence, goal-setting, goal-pursuit – are triggered by the non-conscious priming of concepts.[4] That is, we know that, in many cases, information outside conscious awareness is causally relevant to the actions we perform. In one well-known experiment, subjects in the experimental group were exposed to words, several of which were descriptive of elderly people. Subjects in the control group were exposed to words with no such bias. After exposure, various behaviors in both groups were observed and recorded, including posture and gait, and even performance on subsequent memory tests. In each case, the non-conscious priming of concepts associated with the elderly altered subjects' motor and cognitive behaviors. Primed subjects, compared to controls, tended to walk more slowly or with poorer posture and also tended to remember less well. A wealth of such priming experiments has established that motor and cognitive behaviors are altered in ways we do not notice by mere exposure to words.

Experiments also show that human behavior is affected by a wide range of factors; exposure to words is merely one such factor. Consider, for instance, our remarkable tendency toward mimicry. From infancy through adulthood, we tend to mimic those with whom we are interacting (e.g., gestures, facial expressions, etc.). The absence of mimicry tends to trigger negative affects; the presence of mimicry tends to trigger positive affect and, eventually, trust and even social cohesion. Experiments by Tanya Chartrand and her colleagues, for instance, demonstrate that the degree to which

[4] The concept 'concepts' is hardly adequate, but I will stick with the term "concepts". The relevant psychological units allegedly primed in these experiments are perhaps best described as associative sets or clusters of information, affects, goals, etc. See below.

we like one another person is strongly affected by the degree to which we mimic one another.[5]

But the effects of non-conscious primes extend well beyond the activation of motor and cognitive behaviors. They also cause us to falsely attribute to ourselves knowledge of our reasons for acting. In one recent experiment, subjects in the experimental group read a story about a university student endeavoring to earn money, thereby priming the concept of earning money. After being primed, subjects were asked to choose to play one of two trivia games. One game was about American government; the other, American politics. The covers of both games depicted images relevant to their contents, including pictures of past American presidents, but one game also depicted photos of $1, $10, and $20 bills. The experimenters predicted that subjects primed with the goal of earning money would tend to choose the game with money-related images on its cover, and that is what happened.

However, after making their choices, subjects were given further tasks to perform. Half the primed subjects were told that the trivia game they had selected was challenging; the others were told their game was relatively easy. Subjects were then given eight questions to answer, four from each game. Experimenters made sure that, if a subject had been told that her chosen game was easy, the four questions she received from that game were indeed easy in comparison to the four questions from the other game. Ditto for subjects who had been told that their chosen game was challenging. After answering all eight questions, subjects were asked to tackle a new task. They were asked to choose from among two sets of helpful tips to read. One set of tips was titled "How to Make and

[5] E.g., Lakin and Chartrand (2003).

Save Money"; the other, "How to Successfully Pursue Challenges". Finally, subjects were asked to indicate which of nine possible reasons account for their earlier choice of trivia game. Among the nine reasons were the desire to make money, whether the topic was challenging or easy, whether the topic was interesting, and so on.

The results are interesting, perhaps troubling. As mentioned, subjects primed to make money tended to choose the trivia game with depictions of money on the cover. What is troubling, however, is that the reasons given for their choice of game did not correlate with this initial prime, but instead correlated with associations created *after* their choice had been made. Those who had been told that their game was challenging tended to report that they were attracted to challenging activities and that *that* was why they had chosen the game they chose. Subjects told that their game was easy reported that they did not care for challenging activities and that *that* was their reason for choosing as they did. This clearly is to confabulate "reasons" for one's earlier choice, since the "reasons" were based entirely on information received *after* a choice had been made. Worse, this is to attribute to oneself a confabulated character trait and then project it back into one's recent past.

It is also interesting how subjects chose between the two sets of tips to read. Remember, subjects primed by the money making story tended to choose the trivia game with money on the cover. But, after receiving the further information that their game was challenging or easy – and thus after attributing to themselves the desire for challenging (or for easy) tasks – their choice of tips reflected not the initial prime but the mistaken self-attributions. This is significant. It suggests that the mistaken self-attribution was not innocuous but instead produced substantial downstream affects on the selection and

pursuit of a concrete goal. It shows that, like Gazzaniga's patients, these subjects generated a "reason" for their action on the basis of incomplete conscious information (the efficacy of the money-making prime is something that no subject noticed), then projected their confabulated reason into their own recent past and even engaged in goal-related behavior on the basis of that "reason".[6]

In Defense of (1): The Extended Psychological Framework

That, then, is a glimpse of the findings in social psychology that support the Social Psych Model, but there are additional grounds, including the convergence of findings from other areas in psychology, namely, cognitive psychology and cognitive neuroscience. Consider first recent work on autobiographical memory. Human memory, we now know, comprises several distinct memory systems, including a system dedicated to cultivating and conserving a sense of one's self. According to Martin Conway (2005), autobiographical memory is partly constituted by a Self Memory System (SMS), a workspace in which self-related memories are retrieved, applied to the present situation, then reconsolidated into long-term memory. Of particular importance, the SMS performs its various functions within the constraints of personality-based goals. Features of one's personality tend to bias the retrieval and the reconsolidation of self-related memories. Thus, a person oriented toward intimacy recalls more readily memories oriented toward family, friends, colleagues, etc. A person oriented toward overcoming obstacles or achieving success retrieves more readily memories concerning past successes or failures.

[6] Daniel Kahneman describes us as "associative machines", as animals who quickly and non-consciously evaluate, respond emotionally and physically, and thereby prepare to react to all manner of stimuli. The several experiments he cites all demonstrate the effects of non-conscious primes on our behaviors, emotions, memories, goals, etc.

Conway's many experiments support the general claim that differences in personality-based goals bias the retrieval and re-consolidation of self-related memories.[7]

Conway does not claim that autobiographical memory engages in wholesale fabrication. But he does claim that there is a tension between the drive for accuracy and the drive for coherence in one's view of oneself. In particular, accuracy is at a premium with short term memory, while coherence is paramount in long-term memory. This is a crucial point: the evidence suggests that accuracy is secondary to coherence in the formation and updating of our long-term memories, memories that contribute to our self-image. If so, then your autobiographical memory system, like mine, probably settles for some degree of inaccuracy in order to sustain our image of ourselves as serious scholars or teachers.

Here, then, is the relevance of all this to the Social Psych Model. The memory system most central to our sense of self is likely to amplify and be amplified by confabulated self-related beliefs. Suppose, for instance, I ask you why you selected a particular trivia game and suppose you answer that you did so because it is challenging and you are attracted to challenges. Suppose you say all this even though the actual cause of your choice was a non-conscious prime to make money. Well, if your confabulation goes undetected, you will subsequently have an episodic memory that includes a false explanation of your own action. If autobiographical memory works as Conway claims,

[7] On Conway's view, personality-based goals constitute an immune system in the psychology of the self. These goals achieve this by perpetuating and thus preserving themselves against change by filtering the retrieval and the re-consolidation of autobiographical memories. The general idea here traces back at least to Rapaport (1952), whom Conway cites: memory should be conceived of "not as an ability to revive accurately impressions once obtained, but as the integration of impressions into the whole personality and their revival according to the needs of the whole personality (p. 112–113)."

this faulty memory may be consolidated into your long-term memory. It may come to comprise part of the way you see yourself. If that happens, then this part of your developing self-image may lead to future instances in which your reasons for acting are mistaken. How? Having acquired the false belief that you are attracted to challenges, the next time you are called on to account for your action, this false belief may rise to conscious awareness. If it does, and if the actual causes of your action are not available to consciousness, or if they are ambiguous due to the complexities of the situation, your interpreter may latch on to what happens to be available to consciousness and thereby lead you astray yet again.

Now consider recent work from cognitive neuroscience on the nature of consciousness. Stanislav Dehaene (2001) defends what he calls the "global neuronal workspace hypothesis". Dehaene's larger aim is to show that there are systematic differences between mental states that subjects describe as 'conscious' and states they describe as not conscious. The aim is to show that there are systematic information-processing processes and reproducible neuronal activation patterns that distinguish conscious from non-conscious states.

The core idea is that there exists a neural "workspace" in which the outputs of multiple specialized brain areas connect in a coordinated manner. The workspace includes top-down attentional mechanisms that amplify the outputs of some of the received outputs. To enter conscious awareness, the outputs received from a specialized brain area must meet two general requirements: they must be ongoing, that is, the specialized brain area must be sending a continuous stream of such outputs. In addition, the outputs must be maintained by attentional mechanisms.

The details are intriguing, but two features of Dehaene's model are important for our purposes. First, many non-conscious processes can never enter conscious awareness for the very simple reason that they lack the requisite architectural connections to the neural workspace. The processes that give us depth perception are like this, as are the processes that implement our ability to name the objects in front of us or reach out and pick up a coffee mug. Second, among the processes that can enter conscious awareness, only a proper subset ever do so, because either they are too intermittent or they are overlooked by attentional mechanisms in the workspace. What these two features make clear is that there are neural processes that potentially contribute to the actions we perform that either *cannot* or *do not* rise to conscious awareness.

So, in general, the Social Psych Model integrates with cognitive psychology and cognitive neuroscience in two ways. First, as in the work of Dehaene, the Social Psych Model is correct as a consequence of the structure of our neural capacities. Second, as in the work in Conway, at least some of the mechanisms that implement distinct psychological capacities probably interact with one another and thus increase the confabulation of reasons. This makes thesis (1) difficult to ignore.

In Defense of (2): The Larger Framework

But that is not all. Considerations that support the Social Psych Model include theories that make plausible the claim in thesis (2) that our reason-giving capacities serve vital social functions whether or not they track the truth. This may suggest that the Social

Psych Model is most applicable in situations that bear on social cohesion.[8] I will briefly illustrate with two examples.

Consider work in affective neuroscience. Jaak Panksepp (1998) claims that the mammalian brain comprises seven primary-process emotional systems, six dedicated to social relations. All of Panksepp's systems are neuro-chemical processes implemented in anatomical structures that are homologous across all mammal species. Perhaps the most important is the SEEKING system. SEEKING constitutes us as energized, expectant explorers of the world, as appetitive organisms eager to investigate our environment. Evidence that such a system exists comes from electrical and chemical stimulation experiments of neuro-chemical tracts that run through the lateral hypothalamus. Among the other primary-process systems – those that function in our social attachments – the PANIC system is of particular interest.[9] PANIC enables organisms to extricate themselves from a range of life threats, from drowning or choking to isolation and loss. The function of this system is fulfilled, for example, by the distress calls of some infant mammals. Such calls begin when the infant becomes mobile and can wander away from its mother and they taper off as the child's capacity to fend for itself begins to blossom.[10] Panksepp also views PANIC as the neural system that moves us to be the kind of social animals we are. His hypothesis is that we do not have a distinct system dedicated to social attachment as, for instance, our visual system is dedicated to vision. Instead, the

[8] We thus might revise the Social Psych Model along the following lines: When our conscious, reason-giving capacities, operating in a variety of situations that bear on social cohesion, are unwittingly faced with the absence of causally relevant, non-conscious information, they invent a "reason" using whatever information happens to be accessible.

[9] The five other systems include LUST, CARE, PLAY, RAGE, and FEAR.

[10] Evidence again comes from electrical stimulation experiments. When homologous anatomical structures are stimulated in infant primates, cats, guinea pigs, and more, the same immediate behavior is observed, namely, separation distress vocalizations.

system that helps free us from life-threatening circumstances also causes us to minimize the pain of loneliness and the terror of separation by motivating us to develop social attachments.

Panksepp's discoveries provide especially compelling support for the Social Psych Model, because the many effects of his systems in our psychology are largely concealed from conscious awareness. The outputs of SEEKING and PANIC, for instance, tend not to rise to conscious awareness or, if they do, we tend to conceptualize them as something they are not. For instance, the efficacy of the SEEKING system is something we tend not to notice, since it is part of the very machinery that constitutes our capacities for noticing. The point concerning PANIC is a bit different. We certainly know that some part of our psychology causes us to panic when, for example, we are choking and cannot catch a breath. In addition, we often experience the need for social attachment or the pain of loneliness. But none of those experiences reveals to us that the panic of social isolation is what moves us to bond with others. A similar point applies to Panksepp's other systems: the effects of our affective systems on our agential capacities are, to a significant extent, inaccessible to conscious awareness. They no doubt affect our deliberations and choices while leaving our left-hemispheric interpreters largely in the dark.[11]

[11] The above line of argument may be challenged by appeal to recent theories of "executive" or "cognitive" control. Such theories, it might be claimed, show that I am overplaying the potential for confabulation, since it is evident that we do have at least some capacity for such control. The obvious response to this objection, which I develop elsewhere, is that all models of cognitive control (e.g., Miller and Cohen 2001) include an affective component integral to the coherence of the models. Once we plug Panksepp's affective systems into these models, it becomes clear that the Social Psych Model applies even to the processing of these hypothesized systems. We are susceptible to the confabulation of reasons even when exercising so-called executive control.

Panksepp's systems also support thesis (2). The core claim is that SEEKING disposes us to look for causal intelligibility, including intelligibility regarding our own actions, but sometimes disposes us to infer the existence of causal relations where none exists. In the same way, when SEEKING and PANIC work together to help us render our social world causally intelligible, we are disposed, in the course of seeking such intelligibility, to infer causal relations that are illusory. Here, vital social functions related to cohesion are fulfilled even when our capacities fail to track the truth.[12]

My final large-canvass consideration comes from recent work in evolutionary anthropology. Sarah Hrdy (2009) argues that our capacities for empathy, reciprocity, and mindreading probably spread through the population thousands of years ago, thanks to their economic and social benefits. Particular benefits studied by Hrdy cluster around *cooperative parenting*. The basic thought is that a network of social support made possible by empathy, reciprocity, and mindreading would have made cooperative parenting a highly effective strategy for reproductive success. To illustrate, consider just a single social relation, namely, reciprocity.[13] It is plausible that our reason-giving capacities evolved as part of a social support network that included this capacity. Why? Because if I secure a large stock of food and share it with you when you are in need, you will likely feel a positive desire to help me in the future. In the same way, if I provide emotional support when you suffer loss, you are likely to feel a positive desire to come to my side when I am facing loss. But what are we are doing when we experience and remember emotions in these ways? Hrdy's claim is that we are engaging in a form of

[12] The argument for this line of reasoning, such as it is, is given in Davies (2011). The above is a cryptic statement of the view.
[13] Capacities also studied by primatologists. E.g., de Vaal (2008).

social-emotional bookkeeping. Without noticing it, we are keeping affective account of acts of good will toward each other. Crucially, this kind of bookkeeping plausibly extends to giving reasons for actions. In many instances, especially when good will has not been established or is under threat, giving reasons for our actions helps create or restore it. Our reason-giving capacities, therefore, plausibly evolved with our capacity for reciprocity in so far as it contributed to social cohesion.

Hrdy's hypothesis, if defensible, may help us answer the obvious question: if our reason-giving capacities are so prone to generate false positives, why on earth do we have then? The answer suggested by Hrdy's theory is, as I say, social and economic. But if the evolved functions of our reason-giving capacities are indeed a form of social-emotional bookkeeping based on perceptions of good or ill will, then the accuracy of our reasons will be secondary to the appearance of good will. Sincerity will trump accuracy with respect to cohesion.[14]

Convergence: Darwin's Strategy

In the *Origin of Species*, Darwin dedicates most of his discussion to a series of abductive arguments. In each case, the argument pits special act creationism against evolutionary theory and, in each case, the evidence is shown to favor the latter. That by itself makes for an effective argumentative strategy. But the power of Darwin's overall argument is further enhanced by its scope. The evolutionary hypothesis is superior not just with respect to domestic breeding, but also with respect to embryology, morphology, geology, the distribution of species, classification, and more. With respect to several

[14] I am not referring to feigned sincerity, but sincerity despite false reasons. As Gazzaniga frequently points out, all of his experimental subjects, when asked for their reasons, answered with no hesitation and with utter sincerity, despite knowing the facts of their surgery.

distinct phenomena concerning the nature of life, the evidence supports evolution and tells against creationism.

Darwin's strategy for the study of life is to be emulated in the study of the capacities of living things, including the putative reason-giving capacities of at least one primate species. That is what I am proposing here. The considerations mentioned above – experiments in social and cognitive psychology, and theories in anthropology and neuroscience – are a small sample of the many considerations that converge upon theses (1) and (2). So unless the above considerations converge more strongly upon some other conclusion, one that conflicts with my skepticism, or unless there is a superior argumentative strategy that does not appeal to any such convergence, we should conclude that there is indeed potent scientific evidence that sometimes, when we confidently claim to know our reasons for acting, we are demonstrably mistaken. If so, then we must take seriously the skepticism articulated in theses (3) and (4) above. That, in outline, is my argument for the claim that the Model Penal code is a defective law.

Legal Pragmatism and the Sciences of the Self

If the above line of argument stands then we might have an interesting challenge for any philosophical theory of the law, including legal pragmatism. The sciences of the human self, on my view, reveal a conflict in the very constitution of our psychology that may thwart our best efforts to fulfill the main function of the law. The conflict is as follows. One part of our psychology disposes us to give reasons for actions, where the activity of giving reasons plausibly contributes to social cohesion. If social cohesion is necessary for social justice, then these same psychological dispositions contribute to the

main goal, or one of the main goals, of the law. Yet a different part of our psychology disposes us to understand how things work, including our own minds, and these dispositions lead us to the conclusion that our reason-giving capacities can fulfill their vital social function even in cases where we are ignorant of our real reasons. Our disposition to study ourselves scientifically thus leads us to my skepticism, to the conclusion that we are not justified in claiming to know whether the reasons we give for our actions are genuine reasons. But if we are not justified in claiming to know the genuine reasons for our actions, then the putative justificatory force of our reason-giving capacities is cast into doubt, and that thwarts our best efforts to fulfill the function of the law.

Of course, an adequate theory of adjudication must offer some guidance in cases where precedent does not suffice. When faced with a situation that is genuinely novel or a scientific discovery that calls past practices into question, an adequate theory should tell us how to proceed. Legal pragmatism may appear ideally suited to handle such novelties or discoveries, given its insistence on empirical considerations and its forward-looking orientation. Indeed, Douglas Lind, in an ambitious defense of legal pragmatism, holds up Benjamin Cardozo as a model pragmatist whose legal orientation was inspired by the views of Dewey and James. For Dewey and James, the ultimate goal of human inquiry is what Lind calls an "intellectual satisfaction" accomplished by pursuing a variety of penultimate goals. According to James, the most important penultimate goal is *consistency* – consistency among our beliefs, between our beliefs and our sensations and intuitions, and so on. By analogy, Cardozo claims that the ultimate goal of the law is social justice, accomplished by pursuing a variety of penultimate goals, especially the

goals of *uniformity* and *impartiality*. These penultimate goals, in turn, are accomplished primarily by extending precedents established in the past to present and future cases.

However, as Lind also points out, the practice of extending precedent is defeasible for Cardozo, precisely when we meet with novel situations or unsettling discoveries. In such cases, judges must exercise creative discretion, with the proviso that they not lose sight of the ultimate goal of social justice. Here is a nice excerpt from Lind's discussion:

> Pursuing the virtue of consistency on occasion becomes an exercise in resolving conflicts between competing lines of precedent or reconciling inconsistencies arising when loyalty to the adjudicative criteria of logic, history, custom, and tradition lead in different or equivocal directions. Such unsettling occasions of trouble in the legal environment require courts to exercise their residual power of creative choice. No one criterion of adjudicative practice overrides all others. Still, Cardozo insisted that in the exercise of creative choice, judges must defer to the "final cause of law": social justice. (Lind 2011, p. 69 (ms))

It would be nice to know what "creative choice" entails, though, as Lind points out, that may be something we cannot discern until faced with a specific novelty or discovery that forces us to exercise our creativity.

This Lindian/Cardozian appeal to creative choice is surely fitting for conflicts that plausibly *can* be resolved or reconciled, but I doubt it helps with the present case. If there exist conflicts that derive from the very structure of our psychology, as I claim, then resolution or reconciliation may be a practical impossibility, beyond the reach of any pragmatic goals. It may be an unhappy, brute fact of our psychological architecture that

the pursuit of one goal unavoidably destroys our ability to pursue some other goal. Short of re-engineering our minds, it may be that the more we learn about the workings of our psychology the less we can persist in social practices that depend upon our ignorance of processes involved. The constitutional conflicts in our psychology that lead to my skepticism may be such a case.

If resolution and reconciliation are impossible in such cases, then the penultimate goal of uniformity and the ultimate goal of social justice may be thwarted. But that does not necessitate stultification. We may terminate such cases by turning our back on the conflict involved and arbitrarily privileging one disposition over the other. But that, it seems, is to forsake adherence to any legal theory. A verdict would be reached not by appeal to the goal of social justice or the penultimate goal of uniformity but merely by the expediency of quitting the case. Genuine adjudication – *justified* adjudication – would be impossible. Bringing such cases to an end would be "practical" is an obvious sense of the term, but the justificatory force of legal pragmatism would appear to have vanished.

REFERENCES

Bar-Anan, Yoav, and Timothy Wilson and Ran Hassin 2010 "Inaccurate Self-Knowledge Formation as a Result of Automatic Behavior," *Journal of Experimental Social Psychology* 46: 884-894.

Conway, Martin 2005 "Memory and the Self," *Journal of Memory and Language* 53, 594-628.

Davies, Paul Sheldon 2009 *Subjects of the World: Darwin's Rhetoric and the Study of Agency in Nature*. Chicago, IL: University of Chicago Press.

Davies, Paul Sheldon 2011 "Ancestral Voices in the Mammalian Mind: Philosophical Implications of Jaak Panksepp's Affective Neuroscience," *Neuroscience and Biobehavioral Reviews* 35, 2036-2044.

Dehaene, Stanislas and Lionel Naccache 2001 "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework." *Cognition* 79: 1-37.

Gazzaniga, Michael 2000 "Cerebral Specialization and Interhemispheric Communication: Does the Corpus Callosum Enable the Human Condition?" *Brain* 123, 1293-1326.

Hrdy, Sarah 2009 *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. Cambridge, MA: Belknap Press of Harvard University.

Kahneman, Daniel 2011 *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux

Lakin, Jessica and Tanya Chartrand 2003 "Using Unconscious Mimicry to Create Affiliation and Rapport," *Psychological Science* 14: 334-339.

Lind, Douglas 2011 "The Mismeasurement of Legal Pragmatism," *Washington University Jurisprudence Review* 3.

Miller, Earl and Jonathan Cohen 2001 "An Integrative Theory of Prefrontal Cortex Function," *Annual Review of Neuroscience* 24: 167-202.

Panksepp, Jaak 1998 *Affective Neuroscience: The Foundations of Human and Animal Emotion*. New York, NY: Oxford University Press.

Rapaport, David 1952 *Emotions and Memory*. New York: Science Editions.

de Vaal, Frans 2008 "Putting the Altruism Back Into Altruism: The Evolution of Empathy," *Annual Review of Psychology* 59: 279-300.