

The Illusion of Freedom Evolves

To appear in *Distributed Cognition and the Will*, David Spurrett, Harold Kincaid, Don Ross, Lynn Stephens (eds). MIT Press. January, 2007.

“All theory is against free will; all experience is for it.” --Samuel Johnson.

1. “All Theory is Against Free Will...”

Powerful arguments have been leveled against the concepts of free will and moral responsibility since the Greeks and perhaps earlier. Some—the hard determinists—aim to show that free will is incompatible with determinism, and that determinism is true. Therefore there is no free will. Others, the “no-free-will-either-way-theorists,” agree that determinism is incompatible with free will, but add that *indeterminism*, especially the variety posited by quantum physicists, is also incompatible with free will. Therefore there is no free will. Finally, there are the *a priori* arguments against free will. These arguments conclude that it makes no difference what metaphysical commitments we hold: free will and ultimate moral responsibility are incoherent concepts. Why? Because in order to have free will and ultimate moral responsibility we would have to be *causa sui*, or ‘cause of oneself.’ And it is *logically* impossible to be self-caused in this way. Here, for example, is Nietzsche on the *causa sui*:

The *causa sui* is the best self-contradiction that has been conceived so far; it is a sort of rape and perversion of logic. But the extravagant pride of man has managed to entangle itself profoundly and frightfully with just this nonsense. The desire for ‘freedom of the will’ in the superlative metaphysical sense, which still holds sway, unfortunately, in the minds of the half-educated; the desire to bear the entire and ultimate responsibility for one’s actions oneself, and to absolve God, the world, ancestors, chance, and society involves nothing less than to be precisely this *causa sui* and, with more than Baron Münchhausen’s audacity, to pull oneself up into existence by the hair, out of the swamps of nothingness.¹

¹ Nietzsche, 1992, pp. 218-219. Thanks to Galen Strawson for drawing my attention to this tactful passage.

The conclusion of all of these arguments is that what we do, and the way we are, ultimately comes down to luck—the luck of the nature draw, the nurture draw, the brain-state draw, and perhaps the quantum indeterministic event draw. So while it may be of great pragmatic value to hold people responsible for their actions, and to employ systems of reward and punishment, no one is *truly deserving* of blame or praise for anything.

My aim here is not to argue directly for this conclusion.² I will say only that 2500 years have passed and the reasoning behind it has never been refuted or, in my view, even seriously undermined. Yet the view that we lack free will and moral responsibility is seldom taken seriously. On the contrary, it is often dismissed out of hand, even by those who recognize the force of the arguments behind it. Philosophers who reject God, Cartesian dualism, souls, noumenal selves, and even objective morality, cannot bring themselves to do the same for the concepts of free will and moral responsibility. The question is: why?

2. “...All Experience is for It.”

We *feel* free. When we face a situation in which we have to make a decision, we feel, at that moment, like we can act with deep metaphysical “contra-causal” freedom. We believe ourselves to be in total control of the decision. We feel like a centralized “will” is about to cause the choice; and, if it is a moral choice, we feel that we will be morally responsible for whatever choice we make. Galen Strawson (1986) gives a vivid description of this experience in the following example. Imagine, he writes, that it is Christmas eve and you are going to buy a bottle of scotch with your last twenty dollar bill. Right outside the liquor store, there is a man with an Oxfam cup, or a beggar who is clearly in need. You must decide now whether to spend your money on the scotch, or give to the beggar.

You stop, and it seems quite clear to you — it surely *is* quite clear to you — that it is *entirely up to you* what you do next — in such a way that you will have deep moral responsibility for what you do, whatever you do. The situation is in fact utterly clear: you can put the money in the tin (or give it to the beggar) or you can go in and buy the scotch. You're not only completely, radically free to choose in this situation. You're not free not to choose. That's how it feels.³

Strawson's point is that the *phenomenology* of decision-making leads us to believe that we are radically free and responsible—at least in the immediate moment. Philosophical theories of free will have understandably tried to save this phenomena, or phenomenology, but at the expense of either dodging the central objections raised by anti-free will arguments or by lapsing into a stubborn mysterianism. (Samuel Johnson's other famous quotation on this subject is “I *know* we're free and there's the end on't.”) There is another option, however—one that remains naturalistic but faces the problem squarely. We may accept the soundness of the arguments against free will, but instead of trying to justify our belief in it, we try to explain it away.

3. An Error Theory of Free Will and Moral Responsibility.

Doing this means providing an ‘error theory’ of free will and robust moral responsibility (RMR):⁴ The error theory makes the following claims.

1. We commonly suppose ourselves and others to have the type of free will that would make us RMR for our behavior.
2. RMR does not exist—no one is ever deserving of blame or praise for any action whatsoever.
3. The widespread belief and experience that we have free will and RMR can be ‘explained away,’ that is, accounted for in such a way that does not involve the actual existence of free will and RMR.

² For recent theoretical arguments in support of this conclusion see G. Strawson (1986) and Pereboom (2001).

³ G. Strawson, 1986, p. x.

⁴ By robust moral responsibility I mean the type of responsibility that would make us truly deserving of blame or praise for our actions. I add ‘robust’ in order to distinguish this desert-entailing type of moral responsibility from other uncontroversial varieties—causal responsibility, the capacity to act according to reason, the capacity to form second order volitions, and other types of “compatibilist responsibility.”

Free will skeptics have developed strong arguments for the first two claims, but they have not paid sufficient attention to claim 3. The result is that the skeptical position is dismissed out of hand. However, if we can provide a plausible explanation for why we *mistakenly* believe ourselves to be free and morally responsible, the first two claims gain even greater support.⁵ And the skeptical position can no longer be ignored.

This project has been undertaken by a few others. Spinoza, for example, in the Appendix to Part I of the *Ethics*, offers the following explanation for our belief in free will:

Men believe they that they are free, precisely because they are conscious of their volitions and desires; yet concerning the causes that have determined them to desire and will they have not the faintest idea, because they are ignorant of them.⁶

Charles Darwin presented a similar type of explanation in his Notebooks:

The general delusion about free will [is] obvious—because man has power of action, & he can seldom analyze his motives (originally mostly INSTINCTIVE, & therefore now great effort of reason to discover them...) he thinks they have none.⁷

This simple and rather profound idea may well be part of the correct explanation. We are aware of our desires and volitions and how they cause our behavior. But in most cases we are ignorant of the causes or motives behind the desires and volitions themselves. Thus, as reflective and self-conscious creatures, we have developed this view of free will, this idea that certain volitions *have* no causes or hidden motives—that they derive from *us*, from the *self*, and only there. We believe we are *causa sui* because we don't know what else could have caused our volitions.

Spinoza and Darwin were, in addition, skeptics about moral responsibility. Darwin, for example, writes:

⁵ The *explanandum* here, in other words, is not the existence of free will. Rather, the *explanandum* is our *belief* that we have free will. One explanation might be that we have free will and that leads us to believe that we do. But there are other competing and perhaps more plausible explanations for our belief and these explanations need to be considered as well.

⁶ Spinoza 1982, p. 57

⁷ Darwin, *Notebooks*, p. 608

This view should teach one profound humility, one deserves no credit for anything. (yet one takes it for beauty and good temper), nor ought one to blame others...One must view a wicked man like a sickly one...⁸

But their explanations for why we attribute robust moral responsibility to ourselves and others are derivative of their explanations of why we believe we have free will. Our ignorance of the causes of our volitions leads to the erroneous belief in free will. And the belief in free will leads in turn to the erroneous belief in moral responsibility. What I want to propose is that a large part of the explanation may be the other way around: the belief in robust moral responsibility leads to the belief in free will.⁹ So then the question becomes: what causes the belief in robust moral responsibility? [Slides 1 and 2]

4. Point of Departure: P.F. Strawson and the Reactive Attitudes

P.F. Strawson's magnificent essay "Freedom and Resentment" suggests a way to answer this question. Strawson at the time he wrote this paper was frustrated with the free will debate, with its emphasis on the question of determinism and on metaphysical debates over the meaning of concepts such as "can" and "possibility"—Strawson thought all of this missed the point. He sought instead to locate our commitment to the concepts of free will and RMR within the general framework of human attitudes and emotions—the "reactive attitudes" as he called them. Attitudes like resentment, gratitude, forgiveness, guilt, and love, are, according to Strawson, "given with the fact of human society."¹⁰ The human commitment to these reactive attitudes, Strawson argued, is too deeply rooted to take seriously the idea that a general theoretical conviction (like a belief in the truth of determinism) could cause us to abandon it altogether. It

⁸ Darwin, *Notebooks*, p. 608

⁹ This is not to claim that the explanation offered by Spinoza and Darwin is false. The illusion of free will may have multiple causes (see below). What I want to suggest is that the need to see ourselves and others as RMR may also have been a factor in the evolution of this (illusory) phenomenology.

¹⁰ Strawson, 1982.

is these reactive attitudes, and not any particular metaphysical theory, that ground our attributions of moral responsibility.

Strawson and the free will error theorist agree on two very important claims: (1) that the reactive attitudes are deeply rooted in human psychology, and (2) that the reactive attitudes are fundamentally connected to the widespread belief in free will and moral responsibility. But whereas Strawson sought to use these facts to ground or justify free will and RMR, the error theorist sees them as part of the explanation for why we *mistakenly* believe ourselves to be free and responsible. The reactive attitudes and our proneness to experience them, according to the error theorist, can help to explain why we commit the error in the error theory of free will.

5. The Reactive Attitudes as Adaptations

The next question we need to ask, then, is why the reactive attitudes are so deeply rooted in human psychology. Strawson never once refers to evolutionary theory in his essay, yet every one of the reactive attitudes he describes has an adaptive rationale. Evolutionary theorists since Darwin have argued that certain emotions and attitudes have been naturally selected to motivate behavior that improves social coordination. The emotions are especially valuable for their ability to motivate behavior that goes against our immediate self-interest but that serves long term reproductive or material gains. Frank (1988), for example, argues that certain problems cannot be solved by rational action. To solve these problems—what he calls commitment problems—we have to commit ourselves to behave in ways that prove contrary to our short-term self-interest. Frank then develops what he calls the ‘commitment model,’ which is “shorthand for the notion that seemingly irrational behavior is sometimes explained by emotional predispositions that help

solve commitment problems.”¹¹ Frank argues that emotions like anger, outrage, guilt, and love serve as ‘commitment devices’: psychological mechanisms designed to counteract the allure of immediate self-interest in favor of long term gains.

Frank’s ‘commitment devices’ and Strawson’s reactive attitudes are virtually identical. The reactive attitudes also match up almost perfectly with the emotions that the evolutionary biologist Robert Trivers (1971) hypothesized would be necessary for creating and enforcing reciprocally altruistic behavior in hominids.¹² According to Trivers, attitudes like resentment motivate what he calls ‘moralistic aggression’: retributive acts that are often out of all proportion to the offense committed (but that serve notice not to repeat the offense in the future). The ‘self-reactive attitude’ of guilt serves two purposes: (1) to prevent individuals from engaging in cheating behavior which would harm the individual in the long run; and (2) to motivate cheaters, after the deed is done, to compensate for their behavior so that future reciprocity can be preserved.

Another role for the reactive attitudes has recently been suggested by the Economist Ernst Fehr. Fehr and colleagues (2004, 2002) argue that reciprocal altruism, kin selection, and Frank’s theory, are part of the story but are still insufficient to explain human cooperation in large groups. They argue that a phenomenon called ‘altruistic punishment’ plays an important role in norm enforcement, working to penalize free-riders enough for cooperation strategies to be adaptive. To support this view, Fehr has conducted a number of public good experiments in which subjects can cooperate or defect to various degrees. After a certain number of rounds, cooperators are given the opportunity to punish defectors at a cost to themselves. Cooperators willingly suffer costs in order to punish defectors in these experiments, even when they know

¹¹ Frank, 1988, p. 11.

¹² Trivers, 1971.

that they will never interact with the defectors again. (This is what makes the punishment ‘altruistic’—they will never benefit from inflicting it.) Why do they do this? Fehr argues that negative emotions, emotions like resentment, serve as the proximate mechanisms for this behavior.

Even a reactive attitude as complex as forgiveness has an adaptive function. Individuals who hold long-lasting grudges lose out on too many cooperative opportunities in prisoner dilemma-like situations.¹³ Indeed, in Robert Axelrod’s famous prisoners dilemma tournaments, one of the primary characteristics of successful strategies was ‘forgivingness.’ Tit-for-Tat, the winner of the tournament, punishes (resents?) one defection. But if the defection is followed by cooperation, then all is forgiven—at least until the next defection.

There is good reason to believe, then, that the attitudes and emotions that Strawson linked with the concepts of free will and moral responsibility were selected for their contributions to biological fitness in hominids.¹⁴

6. If Free Will Did Not Exist, We Would Have To Invent It.

Still, one may ask what all of this has to do with the belief in free will and moral responsibility. Frank never mentions such a belief, nor does Trivers. Furthermore, as Franz De Waal and other primatologists have argued, *Chimpanzees* may very well feel a kind of moral outrage; yet we do not attribute a belief in free will and responsibility to *them*. All of this is true. But we differ from Chimpanzees and other intelligent social creatures in a crucial and

¹³ Alexrod, 1984.

¹⁴ Flanagan (2000) is, to my knowledge, the first to explicitly consider (albeit in a different context) whether the reactive attitudes might be biological adaptations. See also Wright (1994) for an interesting discussion of the adaptive value of the moral emotions.

relevant respect: we are able to question the *rationality* of these attitudes and the accompanying behavior.

First, what do I mean by rationality? For the purposes of this argument, I will divide ‘rationality’ into two parts: ‘a-rationality’ and ‘b-rationality.’ To believe something is a-rational is to believe that it is ‘makes sense,’ i.e. that it is consistent with other beliefs that one holds to be true. So, for example, when I was mad at my wife for not sitting in our green rocking chair during the bottom of the ninth inning in game 4 of the Red Sox/Cardinals world series (because that is where she sat in the bottom of the ninth inning in game 7 against the Yankees), I was being a-irrational. For my anger contradicted a belief I hold to be true: namely, that the actions of two people in Hillsborough, NC have no causal effect on a baseball game being played in St. Louis (especially with the 5 second satellite delay). The second sense of rationality I employ, ‘b-rationality’, is instrumental. To believe that something is ‘b-rational’ is to believe that it serves our short-term material self-interests. So if my goal is to get to class on time, and driving will be faster than walking, then taking the car would be b-rational. (And walking would be b-irrational.)

Let us call creatures who have the capacity to question the a-rationality and b-rationality of their emotions ‘cognitively sophisticated’ or CS creatures. My central contention in this essay is this: for CS individuals, the commitment devices, or reactive attitudes, cannot work as effectively to motivate adaptive behavior unless they are accompanied by a belief in free will and RMR.

Why is that? Well, let us first consider creatures who lack this capacity, who cannot question the rationality of their attitudes. De Waal (2000), for example, tells a story of chimpanzee indignation or outrage and the retributive behavior it motivates. One of the female

chimpanzees De Waal studied, Puist, had earlier supported a male, Luit, against his rival Nikkie—she had helped chase Nikkie away. Later when Nikkie displayed at Puist, an act of aggression, Puist turned to Luit in search of support. But Luit did nothing to protect her. This so infuriated Puist that rather than attack Nikkie, the aggressor, she turned on Luit, barking, and chasing him across the enclosure, even hitting him.

De Waal interprets Puist's behavior as follows. She experienced a type of indignation or outrage at Luit for breaking a chimpanzee social norm, one that simply states "one good turn deserves another." This indignation in turn motivated her to punish Luit, even at risk to her own personal safety. Luit will think twice about breaking that norm again in the future, as will any other chimpanzee who witnessed the incident. But if Chimpanzees are able to experience indignation or outrage, they are certainly incapable of assessing whether or not the indignation is rational. Puist was not capable of asking herself whether the attitude and the accompanying behavior 'made sense,' nor in any sophisticated way, whether they served her self-interest. The attitude did all the motivating work, with no backtalk or interference from the reasoning faculty.¹⁵

[Slide 3]

Now consider a similar scenario involving a CS individual. Luke, a hunter-gatherer, has been wronged by Gus. Gus did not defend Luke when he should have. (Or maybe he tried to free-ride during the hunt, or attempted to steal Luke's mate.) Luke gets a visceral feeling of indignation or outrage, "mostly instinctive" as Darwin says, and the feeling motivates him to act with moralistic aggression the next time he sees Gus. If Luke were a non-CS creature, that would be the end of the story. The indignation would motivate the retributive behavior, and

Luke would act. And sometimes no doubt that will still happen. But Luke, unlike Puist the Chimpanzee, can question the rationality of this attitude. He can recognize that Gus is larger than he is, and that he may well get the short end of any attempt at revenge. He can also realize that what's done is done, that there is little point in dwelling on something that already happened. Yes, he has learned to stay away from Gus, and not to trust him. But there is no point in risking his own safety, perhaps his own life, to punish him. These rational considerations may well undermine the link between attitude and behavior, and if Gus knows this, he will be more likely to commit the offense. For the risk is lower. That is the essence of a commitment problem.

[Slide 4]

Now I'm not suggesting that early hominids were reflective enough to have all of these thoughts, or to think of them in such a collected manner. But it seems plausible that considerations like these might have undermined the link to some degree between the commitment devices, the reactive attitudes, and the accompanying behavior. If this is correct, then a new design problem arises for CS creatures. The reactive attitudes or the commitment devices are still adaptive for their role in solving commitment and coordination problems. But greater cognitive sophistication has diminished the capacity of these emotions to motivate the accompanying adaptive behavior. CS creatures, then, need something else to offset the dampening effect of increased cognitive sophistication. This 'something else,' I suggest, is an independent belief in the robust moral responsibility of other agents.

How would this work? Well, what if, in addition to the visceral anger, Luke also had a belief or feeling that Gus *deserved* blame, that he *ought* to be punished for what he did? What if Luke felt that, his own interests aside, something would be deeply wrong with the world if Gus

¹⁵ You do not have to agree with De Waal's interpretation of Puist's behavior to appreciate my point here. If you do not believe that Chimpanzees can feel moral outrage, then simply imagine an early hominid reacting to a situation

got away with the offense? Now the outrage is no longer a-irrational. It makes perfect sense. Luke is outraged because Gus deserves blame and punishment for his offense. And though it still might be b-irrational, contrary his to self-interest, to act retributively, this consideration might be offset by that belief that something would be deeply wrong if Gus's offense went unpunished. This belief would then fortify the link between the outrage and the retributive act.

[Slide 5]

We can see this belief in RMR in action by referring to an example Robert Frank uses to illustrate his commitment model. Jones has a \$200 leather briefcase that Smith wants to steal. If Smith steals it, Jones will have to go to court to recover it and force Smith go to jail for 60 days. But the day in court will cost Jones \$300 in lost earnings not to mention the tediousness of a court trial. Since this is more than the briefcase is worth, it is clearly not in Jones' material self-interest to press charges. The problem, of course, is that if Smith *knows* that Jones is going to be rational in this way, then he can steal the briefcase with impunity. There's no risk. *But*, Frank writes:

Suppose that Jones is *not* a pure rationalist; that if Smith steals his briefcase, he will become outraged, and think nothing of losing a day's earnings, or even a week's, in order to see justice done. If Smith knows Jones will be driven by emotion, not reason, he will let the briefcase be. If people *expect* us to respond irrationally to the theft of our property, we will seldom *need* to, because it will not be in their interests to steal it. Being predisposed to respond *irrationally* serves much better here than being guided only by material self-interest.¹⁶

From this passage, one might think that the outrage alone is sufficient to predispose Jones to act 'irrationally.' Frank makes no explicit reference here to any beliefs about moral responsibility or free will. But the belief is there—implicit in the remark about Jones' need to see "justice done." If Jones did not believe that Smith *deserved* blame and punishment for stealing the briefcase,

similar to the one De Waal describes.

then he would feel no need to see justice done. And the outrage, no matter how fierce, might very well not be enough. Suppose Jones viewed Smith like he viewed a dog. He might be angry and frustrated about losing the briefcase, but he would feel no burning need to sacrifice his own interests to pursue and punish the dog who stole his briefcase. Without the belief, then, that other human beings deserve blame or punishment for their actions, Jones's outrage would have been insufficient to solve this commitment problem. With the belief, the link between attitude and adaptive behavior is fortified.

So where does the experience of free will come into the picture? Well taking this account even further, we may say that the phenomenology of free choice is a very complex adaptation that allows us to attribute robust moral responsibility to ourselves and others. This part of the story owes what Shaun Nichols has termed, in a slightly different context, "a perverse debt to Kant."¹⁷ For to paraphrase Kant, the moral life of *rational creatures* cannot get off the ground unless we experience ourselves as indeterministically free agents. If we did not believe that we could have acted otherwise, or intended otherwise, or been other than we are, then how could we see ourselves as RMR for an action? And if we did not believe that others were indeterministically free agents, then how could we see offenders as deserving of blame or praise for *their* behavior? If Kant is right that libertarian freedom is a necessary condition for RMR, then possessing the illusion of free will would allow us to believe that attributing RMR to ourselves and others was a-rational. And if, as I have argued, the attribution of RMR is adaptive for CS creatures, then the phenomenology of free will, the illusion that we can act freely and be ultimately responsible for our actions, would have been adaptive as well.

To recap, my argument makes the following claims.

¹⁶ Frank, 1988, p. x

¹⁷ Nichols, 2004. Perverse, because neither he nor I endorse Kant's (libertarian) conclusions.

1. Retaining the link between the reactive attitudes and the accompanying behavior was important for biological fitness in hominids.
2. Greater cognitive sophistication, which evolved for other reasons, undermined this link. The ability to assess the a-rationality and b-rationality of our attitudes made it less probable that the attitude would motivate the accompanying adaptive behavior.
3. A belief in the robust moral responsibility of oneself and others would offset this undermining effect, and would therefore have been adaptive.
4. However, CS creatures do not consider it a-rational to attribute robust moral responsibility to agents who are not radically free.
5. Therefore, unless we believed ourselves and others to be radically free agents, we would not consider it rational to attribute RMR.
6. Experiencing ourselves as radically free agents allows us to believe that we and others are radically free.
7. Therefore, the experience of radical free agency, the phenomenology of free choice, the illusion of the conscious will, would have been adaptive for CS creatures.

The conclusion, I have to stress, is not that the sole cause of the illusion of free will is the need to attribute RMR to ourselves and others. (Nor am I in any way claiming that cultural forces and social structures have had no effect on these concepts as we understand them today. Surely they have.) The phenomenology of free will may have multiple causes and origins, and the advantages outlined in my hypothesis may be only a small part of the explanation for why we possess it.¹⁸ Don Ross's theory, in this volume, that selves arise as stabilizing devices for social

¹⁸ Furthermore, the illusory phenomenology surely did not come in one fell swoop after the reactive attitudes were in place. Thanks to Andy Clark for pointing out during the presentation of this paper how the theory could be interpreted in this (highly implausible) manner. If my account is correct, or approximately correct, there must have

dynamics is a highly plausible—and I believe, complementary—explanation for why we experience ourselves as having an autonomous, if predictable, self. Consider Ross’ claim that “the massive interdependency among people incentivizes everyone to regulate the stability of those around them through dispensation of *social rewards and punishments*.” [my italics]¹⁹ My hypothesis is that the belief in RMR enables this dispensation of social rewards and punishment to occur more effectively. And again, perhaps Spinoza and Darwin are right as well: the origin of the illusion of free will is that we are aware of the desires and volitions that cause our actions, but unaware of what causes the desires and volitions themselves. My theory, then, would point out the adaptive advantages of having this phenomenology once it is in place. And the belief in free will would arise from this phenomenology *and* the need to justify our belief in RMR.

Another point I wish to stress is this. The “explaining away” of RMR should not be taken to support the claim we lack other important (non-desert-related) varieties of responsibility. As compatibilists from Hume to Frankfurt to Dennett to Pettit have observed, there are numerous valuable types of responsibility that are entirely compatible with the most uncompromising naturalism imaginable. The error theorist does not disagree, for example, with Philip Pettit’s claim that we are “conversable”—that is, that we have the ability to track and conform to relevant reasons given to us in conversation—or that this capacity is fully compatible with determinism.²⁰ Rather, the error theorist and the compatibilist disagree on the question of whether this capacity can truly justify robust desert-entailing moral responsibility. Elsewhere²¹ I have argued that compatibilist forms of freedom and responsibility are insufficient for this task. The argument presented in this paper supports this claim at best indirectly—by showing why

been a co-evolution of sorts between the phenomenology of free will and the increasing complexity of the reactive attitudes.

¹⁹ See Ross, this volume.

²⁰ See Pettit, this volume.

people (including philosophers) are so resistant to give up the concept of robust moral responsibility.

7. Testing the Hypothesis.

To fend off some inevitable “just-so story” objections, I want to highlight three claims I’ve made that are testable.

(1) *Human beings (CS creatures) who deem an attitude, in a certain situation, to be irrational are less likely to perform the behavior that is often motivated by this attitude.* This is a key claim but one that seems amenable to fairly straightforward empirical investigation.

Experiments in social psychology might help to determine the degree to which our behavior is governed by what we deem to be rational or irrational, and especially what occurs when we think an *attitude or emotion* is irrational.

(2) *We believe we act with libertarian (indeterministic) free will.*

(3) *We do not believe it rational to attribute RMR to an agent unless we believe they have libertarian free will.*

These are essentially questions about what our intuitions are regarding free will and moral responsibility. Are they compatibilistic or incompatibilistic? Do we really think an agent needs to have acted with indeterministic free will in order to believe that our assignments of blame are rational and justifiable? Until recently, pronouncements on this question were made from the philosopher’s armchair. Now, however, philosophers and psychologists are conducting studies that will help determine whether our intuitions are compatibilist or incompatibilist

²¹ Sommers, 2005

regarding our own actions, and about the actions of others.²² Much more should and will be done on this front.

In addition, more can be said on the more general conditions required for attributing blameworthiness. Malle and Nelson (2003), for example, contend that in order to hold someone blameworthy, we must believe they acted intentionally. But sometimes we assign blame *first*, and then apply the concept of intentionality to the behavior afterwards, whether it fits or not, in order to justify the blame already assigned [Knobe, forthcoming]. This model is quite congenial to the error theory I've described, for it shows how the assignment of moral responsibility may come first, and the explanations for why agents are responsible come afterwards (even if the explanations are false). My contention is that in addition to coming first in the psychological process, assignments of blame may have come first in the evolutionary process as well. The phenomenology of free will and RMR emerged as a means of justifying blame-linked attitudes like resentment.

9. Conclusion.

Of course this account, even if true, does not on its own demonstrate the illusoriness of free will and RMR. But a plausible 'explaining away' of free will, when *combined* with the powerful negative arguments against free will and RMR makes a good case for an error theory. I have argued that Darwinian theory can be of great help in providing this explanation.

It's important to note, however, that this evolutionary approach to explaining our commitment to the concept of free will and RMR differs crucially from other theories (e.g. Dennett, 2003) that attempt to "naturalize" freedom and responsibility. For whereas I agree that

²² See for example: Nahmias et al (2005), Nichols (2004), Nichols and Knobe (unpublished manuscript). The data are, in my view, insufficient for any side to claim victory at present.

the experience of free will and RMR can be explained naturalistically, I am nevertheless convinced that it *is* an illusion—that true moral responsibility, and the retributive attitudes that presuppose it, are unjustifiable. I am also convinced that the unraveling of the illusion would have significant ethical, legal, and practical implications, both for the individual, and for a society that embraced an error theory of free will.

Regarding these implications, the prevailing view among philosophers—and non-philosophers for that matter—is that they would be disastrous. Everything would be permitted. Life would lose most or all of its meaning; we would be puppets on a string, living a mockery of a real human existence. As I have argued elsewhere, however, this pessimism is little more than a contemporary prejudice—it is seldom argued for, and almost always based on a distorted view of what free will skepticism really entails.²³ A denier of free will and RMR may live a happy, fulfilling, love-filled, moral life without in any way contradicting his or her principles. Indeed, one may say, as Darwin did about his theory of natural selection that there is grandeur to this view of life. Or as Einstein has written: “This realization [that there is no such thing as free will]...prevents us from taking ourselves and other people too seriously; it is conducive to a view of life which, in particular, gives humor its due.”²⁴

Acknowledgments

I am grateful to Alex Rosenberg, Owen Flanagan, Eddy Nahmias, Shaun Nichols, Joshua Knobe, and Manuel Vargas for comments on earlier drafts of this paper. I have also profited immensely from discussions at the *Mind and World* conference at the University of Alabama at Birmingham.

²³ Sommers, 2005.

²⁴ Einstein, 1984, pp. 8-9

References:

- Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.
- Darwin, C. *Charles Darwin's Notebooks, 1836-1844*, Cornell University Press.
- Dennett, D. (2003). *Freedom Evolves*. Viking.
- De Waal, F. (2000). *Chimpanzee Politics*. Johns Hopkins University Press.
- Double, R. (1996). *Metaphilosophy and Free Will*. Oxford University Press.
- Einstein, A. (1982). *Ideas and Opinions*. Crown Publishers.
- Fehr, E. (2004). "The Neural Basis of Altruistic Punishment." *Science*. 305: 1254-1258.
- Fehr, E. and Gächter, S. (2002). "Altruistic Punishment in Humans." *Nature*. 415: 137-140.
- Flanagan, O. (2000). "Destructive Emotions." In *Consciousness and Emotion* 1:2.
- Frank, R. (1988). *Passions Within Reason*. Norton.
- Knobe, J. (Forthcoming). "The Concept of Intentional Action: Case Studies in the Uses of Folk Psychology." *Philosophical Studies*. Forthcoming.
- Malle, B. F., & Nelson, S. E. (2003). "Judging mens rea: The tension between folk concepts and legal concepts of intentionality." *Behavioral Sciences and the Law*, 21: 563-580.
- Nahmias, E. et al. (2005). "Surveying Free will: Folk Intuitions about Free Will and Moral Responsibility." Forthcoming in *Philosophical Psychology*.
- Nietzsche, F. (1992). *Beyond Good and Evil*. In *The Basic Writings of Nietzsche*, edited by Walter Kaufmann. The Modern Library.
- Nichols, S. (2004). "The Folk Psychology of Free Will: Fits and Starts." *Mind and Language*. 19: 473-502.
- Nichols, S. and Knobe, J. "Moral Responsibility and Determinism: Empirical Investigations of Folk Intuitions." Unpublished Draft.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge University Press.
- Pettit, Philip. "Neuroscience and Agent-control" This volume.
- Ross, D. "The Economic and Evolutionary Basis of Selves." This volume.
- Sommers, T. (2005). *Beyond Freedom and Resentment: An Error Theory of Free Will and Moral Responsibility*. Dissertation, Duke University.
- Spinoza, B. (1982) *The Ethics and Selected Letters*. Edited by Seymour Feldman. Hacking Publishing Company
- Strawson, G. (1986). *Freedom and Belief*. Oxford University Press.

Strawson, P.F. (1982) "Freedom and Resentment." From *Free Will*. Edited by Gary Watson. Oxford University Press.

Trivers, R. (1971). "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology*. 46:35-56.

Wright, R. (1994). *The Moral Animal*. Vintage.