Ethics in Progress (ISSN 2084-9257). Vol. 5 (2014). No. 2. 05-20.
doi: 10.14746/eip.2014.2.2

Free Will and Retribution Today

Mario De Caro (Università Roma Tre and Tufts University, Roma)
Massimo Marraffa (Università Roma Tre, Roma)

Shaun Nichols has proposed a useful distinction regarding three different projects in the inquiry of free will and responsibility: a descriptive project, a substantive projects, and a prescriptive project[1]. In this article we address two issues that have been recently debated in the literature on free will, moral responsibility and the theory of punishment: the first issue concerns the descriptive project, the second both the substantive and the prescriptive project.

The first issue concerns the impact that the evidence for determinism, supposedly shown by cognitive sciences, would have, if popularized, on the ordinary practice of responsibility attributions. On theoretical, historical and empirical grounds, we claim that there is no rationale for fearing that the spread of neurocognitive findings will undermine the folk practice of responsibility attributions.

The second issue concerns the consequences that a demonstration of the illusoriness of moral responsibility would have for the theory of punishment. In this regard, two opposite views are advocated: (i) that such a demonstration would cause the collapse of all punitive practices; (ii) that, on the contrary, such a demonstration would open the way to more humane forms of punishment, which would be justified on purely utilitarian grounds. We will argue that these views are both wrong, since whereas a sound punitive system can be justified without any reference to moral responsibility, it will certainly not improve the humaneness of punishment.

1. Does the Belief in Determinism Prevent Us from Ascribing Responsibility?

Would the dissemination of the findings regarding the neurocognitive bases of human agency undermine our ordinary practice of responsibility attributions? According to Saul Smilansky, it would. With his "free will illusionism," he assumes that the majority of people have illusory beliefs concerning the existence of libertarian free will. He then suggests that if people were

[1] The descriptive project aims "to determine the character of folk intuitions surrounding agency and responsibility." The substantive project makes the effort "to determine whether the folk views are correct." The prescriptive project asks the question "whether, given what we know about our concepts and the world, we should revise or preserve our practices that presuppose moral responsibility, like practices of blame, praise, and retributive punishment" (Nichols 2006, 58-9).

disillusioned about this—that is, if they came to realize that libertarian free will is both incoherent and non-existent—this would lead to catastrophic personal and societal consequences. People would no longer find meaning and value in their lives, and would be less likely to behave morally. Consequently, it would be preferable that philosophers and scientists who know the (horrible) truth about the nonexistence of libertarian free will conceal it in order to avoid moral nihilism. As Smilansky puts it, "humanity is fortunately deceived on the free will issue, and this seems to be a condition of civilized morality and personal value" (2002, 500), and "there would be considerable room for worry if people became aware of the absence of libertarian free will" (2000, 505, note 7).

In our view, however, Smilansky is wrong. On theoretical, historical, and empirical grounds one should doubt that the loss of faith in our free will would have bleak implications at both the societal and personal levels. From a historical point of view it is useful to look at the periods in which, generally for religious reasons, there have been communities convinced that free will is illusory. An instructive example, in particular, is offered by the Lutheran and Calvinist communities of the first decades, which accepted the ideas of their respective founding fathers on bound will and predestination. In the sixteenth century, of course, these views were not based (as they are today) on the idea that the natural world is governed by deterministic laws, but by God's prescience and providence. From the point of view that interests us here, however, this fact does not change the substance of the phenomenon: what is interesting to notice, in fact, it is that the members of those communities were convinced that, since free will was denied to human beings, they were not in control of their choices, deeds and lives, and consequently that they were not responsible for what they did. Therefore, on a religious ground, they held with full awareness the beliefs in the illusoriness of free will; moreover, religion mattered very much for those communities. If Smilansky were right, therefore, very bleak consequences should have followed at the social, judiciary and political levels—but this, as is well known, did not happen at all. On the contrary, according to the classic analysis offered by Max Weber in *The Protestant Ethic and the Spirit of Capitalism*, it was exactly because of the certainty with which the protestant communities of the origins—and especially the Calvinist ones (for which the idea of predestination was all-reaching)—believed humans do *not* enjoy free will that they found an extraordinary energy in their worldly acting: to the point that, according to Weber, the product of that attitude was the birth of capitalism. Analogously, if one considers other communities that accept that free will is an illusion, one finds that from this belief no destructive social consequences follow. On the contrary, sometimes, contra the case of sixteenth-century Protestantism, some forms of fatalism followed, but in these cases the final product was social quietism—which, *pace* Smilansky, certainly cannot be seen as a form of widespread social disorder.

But, besides this historical analogical argument, one can appeal also to theoretical reasons for claiming that the inference from the awareness of being unfree does not imply that we should abandon the beliefs in moral responsibility and the retributivist conception of punishment. Along this line, in the already mentioned *Freedom and Resentment*, P. F. Strawson developed two influential arguments, which have been defined "rationalistic strategy" and "naturalistic strategy"[2]. According to the rationalistic strategy, even if we were convinced that determinism is true, it would be rational for us to keep maintaining the system of reactive attitudes and attributions of responsibility on which our view of agency and personhood is presently based. In other words, human beings should keep looking at themselves as agents, not as mere natural objects (as it happens when we deal with little children or with the mentally handicapped); this is because conceiving human beings *only* as natural objects would be irrational since that attitude would cause us to diminish the value of our lives. According to the naturalistic strategy, instead, it can be argued that it is a "natural fact" about human beings that they could never be able to abandon the system of reactive attitudes and attributions of responsibility – whatever science or philosophy has to say on the subject.

In this light, even if we reached theoretical certainty that we are causally determined and that incompatibilism is the correct view of free will (incompatibilism is the view that determinism and free will are incompatible) this fact would have no practical consequences for our lives: we would keep interacting in the usual ways, by considering each other responsible for our respective deliberations and intentional actions (with the usual exceptions concerning mental pathologies and little children). Therefore the idea of responsibility, and all connected practices (including the legal ones), would always be unaffected by our philosophical beliefs.

Moreover, the claim that a belief in determinism would have no bleak implications, can be argued not only on theoretical and strictly philosophical grounds, but also at an empirical level. It is well known that philosophers have made competing claims about people's intuitions about freedom and responsibility. And the received view is Smilansky's one: most people are "intuitive incompatibilists"[3]. But typically these claims have been made with speculative-apriori analysis about what our folk intuitions on the issue are. Recently, however, some experimental philosophers have tested these claims about folk judgments of free will and moral responsibility in a more

[2] We are following Russell's (1995) useful analysis here.

[3] Robert Kane, e.g., writes: "In my experience, most ordinary persons start out as natural incompatibilists. They believe there is some kind of conflict between freedom and determinism; and the idea that freedom and responsibility might be compatible with determinism looks to them at first like a »quagmire« of evasion (James) or »a wretched subterfuge« (Kant). Ordinary persons have to be talked out of this natural incompatibilism by the clever arguments of philosophers" (1999, 217). And Galen Strawson: "It is in our nature to take determinism to pose a serious problem for our notions of responsibility and freedom" (1986, 89).

systematic way[4]. The results have been mixed: it is not clear yet if laypersons are naturally incompatibilist or compatibilist.

Nahmias and colleagues, for example, have carried out a number of studies in which participants (college students who were unaware of the free will debate) were shown three different scenarios describing deterministic universes[5]. Following each scenario, participants were asked a range of questions, including whether a certain person in that scenario acted freely and was morally blameworthy. One of these scenarios was the "Jeremy case":

> Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25th, 2150 A.D., twenty years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00 PM on January 26th, 2195. As always, the supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 PM on January 26th, 2195 (Nahmias *et al*. 2005, 566).

Subjects were asked to indicate whether or not Jeremy robbed Fidelity Bank of his own free will; and whether or not he was morally blameworthy for robbing the bank. A large majority of the subjects gave *compatibilist* answers to both questions: 76% judged that Jeremy acted of his own free will; 83% responded that Jeremy was morally blameworthy for robbing the bank[6]. These results seem to offer evidence that a significant majority of laypersons are "natural compatibilists", and thus call for an explanation for why so many philosophers —including Smilansky—have, on the contrary, thought that most people have intuitions that support incompatibilism.

According to Nichols and Knobe (2007), the divergence between Nahmias et al.'s psychological findings and philosophers' claims, stems from the instability of our intuitions on vignettes of freedom, determinism, and responsibility. Their hypothesis is that the participants are more likely to give compatibilist answers to *concrete* questions about particular affect-laden cases (such as robbing a bank or killing a man), but incompatibilist answers to *abstract* questions concerning more general moral principles. If so, the

[4] Experimental philosophy is a new area of research that involves the gathering of empirical data to tackle philosophical problems. Cf. Knobe & Nichols (2008).

[5] For complete information on the methodology and results of these studies, see Nahmias, Morris, Nadelhoffer, & Turner (2005). For further discussion of the philosophical implications of these studies and this methodology, cf. Nahmias, Morris, Nadelhoffer, & Turner (2006).

[6] For similar findings, cf. Woolfolk, Doris, & Darley (2006).

difference between psychological findings and philosophers' claims is due to a difference between two different ways of *framing* the relevant question.

In one experiment, participants were presented with a description of two universes, A and B. Universe A is a universe alternate to ours, in which "everything that happens is completely caused by whatever happened before it." This includes human decisions, which participants are told "had to happen" as they did. By contrast, in Universe B "almost everything that happens is completely caused by whatever happened before it"; the one exception is human decision making, and hence decisions do not have to happen in the way in which in fact they do happen. When asked which universe more closely resembles our own, 90% of subjects chose the indeterministic Universe B.

Participants were then randomly assigned to one of two groups, one of which was presented with a scenario in the "abstract" condition, and the other in the "concrete" condition. The subjects in the abstract condition were asked the following low-affect question: "In Universe A, is it possible for a person to be fully morally responsible for their actions?" In this condition, 86% of subjects gave the incompatibilist response that it is not possible for a person to be fully morally responsible in Universe A. By contrast, in the concrete condition participants were asked a high-affect question:

> In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family. Is Bill fully morally responsible for killing his wife and children? (Nichols & Knobe 2007, 110-111).

In this concrete and high-affect condition, 72% of subjects gave the *compatibilist* response that Bill is fully morally responsible for killing his wife and children.

According to Nichols and Knobe, these findings suggest that the intuitions about the relations between determinism and responsibility vary according to the affective framing of the scenario. When participants come to deal with macroscopic violations of moral norms, they experience a "reactive attitude" (e.g., moral anger or indignation) that makes them unable to correctly apply the underlying *incompatibilist* folk theory of moral responsibility. The compatibilist intuitions are then the result of "performance errors" caused by the distorting influence of emotion on moral judgment; that is, the bias triggered by the strong emotion prevents subjects from making the inference that by contrast works at the abstract level and leads to conclude that determinism rules out responsibility. This hypothesis is called "Performance Error Model," and it claims that folks have *apparent*, but not *genuine* compatibilist intuitions.

However, Nahmias and collaborators (Nahmias 2006; Turner & Nahmias 2006; Nahmias, Coates, & Kvaran 2007; Nahmias & Murray, forthcoming) have

argued that Nichols and Knobe's scenarios do not allow us to interpret their evidence as supporting the claim that most people have incompatibilist intuitions. Nahmias postulates that what induces the participants in Nichols and Knobe's experiment to deny free will and moral responsibility is their presupposition that determinism implies that the causes of behavior *bypass* our deliberations and conscious purposes. If the "bypassing" construal of determinism were correct, compatibilism would imply the truth of *epiphenomenalism* about our conscious mental life, or it might take the form of *fatalism*. However, Nahmias argues, to think that determinism entails such a "bypassing" is simply a mistake. According to compatibilists, determinism does not make our deliberations and conscious purposes causally irrelevant to what we do. For so long as our mental states are part of a deterministic sequence of events, they play a crucial role in determining what will happen.[7] But "if the reason people express incompatibilist intuitions is that they mistakenly take determinism to entail bypassing, then those intuitions do not in fact support the conclusion that determinism, when properly understood, is incompatible with free will" (Nahmias & Murray, forthcoming).

Summarizing, whether laypersons are naturally incompatibilist or compatibilist is still an open empirical question. This said, what is particularly interesting for us here is that a study by Roskies and Nichols (2008) has put in relation the experimental philosophy of free will with Smilansky's prediction that the acceptance of determinism will lead to moral catastrophe. Roskies and Nichols have explored a important difference in the aforementioned studies: In Nahmias and colleagues' "Jeremy case" the scenario is set in *our own* world, whereas in Nichols and Knobe's study the scenario is set in an *alternate* universe. Roskies and Nichols predicted that intuitions about free will and moral responsibility would be sensitive to whether deterministic scenarios are described as actual, in our world, or merely possible (i.e., true in some other possible world).

To test this prediction, a group of subjects were randomly assigned to one of two conditions: Actual and Alternate. In both conditions participants were shown two short non-technical descriptions of a deterministic universe. In the Actual condition, the description clearly implies that one is talking about our universe:

> Many eminent scientists have become convinced that every decision a person makes is completely caused by what happened before the decision—given the past, each decision *has to happen* the way that it does. These scientists think that a person's decision is always an

[7] But then it is not determinism but reductionism that threatens the notions of free will and responsibility: "… a principal psychological mechanism that drives incompatibilist intuitions involves people's fear of reductionistic descriptions of deliberation and decision-making" (Nahmias 2006, 229). Roskies (2006, 422) notes that similar points have been made by Kim (1998), Flanagan (2002), and Dennett (2003).

> inevitable result of their genetic makeup combined with environmental influences. So if a person decides to commit a crime, this can always be explained as a result of past influences. Any individual who had the same genetic makeup and the same environmental influences would have decided exactly the same thing. This is because a person's decision is always completely caused by what happened in the past (Roskies & Nichols 2008, 3).

In the Alternate condition, it is explicitly said that the universe is not ours by adding to the just mentioned passage the following incipit: "Imagine an alternate universe, Universe A, that is much like earth. But in Universe A, many eminent scientists have become convinced that in their universe… ."

In both conditions, participants were then asked to rate their level of agreement with three statements: in such a world:

(i) it is impossible for people to be fully morally responsible for their actions;

(ii) people should still be morally blamed for committing crimes; and

(iii) it is impossible for people to make truly free choices.

The results were that, compared with the participants in the Actual condition, the participants in the Alternate condition gave higher ratings of agreement to the statements (i) and (iii) but lower levels of agreement to the statement (ii). In short, when participants were asked to imagine that *our* universe was deterministic, they tended to answer that in such a situation it would still be possible for agents to be morally responsible, blameworthy, and able to make free choices (i.e., they would give responses in accordance with compatibilism). However, participants were likely to say that people in an *another* deterministic universe would not be fully responsible and not able to make free choices, and hence that perpetrators of crimes would not be blameworthy. And the latter is of course an incompatibilist response.

In this light we can therefore reconsider our issue: would a widespread acceptance of determinism undermine the folk practice of responsibility attribution? Roskies and Nichols argue that their experiment is relevant to this issue in at least two ways. First, it puts forward a hypothesis about the source of the widespread expectation that the belief in determinism would lead us to give up our belief in moral responsibility: when the deterministic world we are asked to consider is not our own (the Alternate case), we tend to claim that moral responsibility is impossible. Second, the experiment suggests that even if we should come to believe that determinism holds in the actual world, such a belief would not undermine our responsibility judgements. This is evidence against Smilansky's claim that a widespread acceptance of determinism would lead people to moral nihilism.

An oft-cited study by Vohs and Schooler (2008), however, seems to offer evidence against Roskies and Nichols' anti-nihilist conclusion. They report two experiments that suggest that, when participants were primed to question

their belief in free will, they were more likely to cheat. In the first experiment they cheated on a cognitive task, and in the second experiment they overpaid themselves for performances on a cognitive task. This study inspired another paper by Baumeister, Masicampo, and DeWall (2009). Using similar methodology, they report that inducing disbelief in free will in participants increases aggression and reduces helpfulness.

Let us consider Vohs and Schooler's first experiment. 30 undergraduates were randomly assigned to two conditions. In the "anti-free-will condition", participants read a passage from a chapter of Francis Crick's *The Astonishing Hypothesis* in which it is claimed that "rational, high-minded people—including, according to Crick, most scientists – now recognize that actual free will is an illusion, and also [claim] that the idea of free will is a side effect of the architecture of the mind" (Vohs & Schooler 2008, 50). In the control condition, participants read another excerpt from Crick's book, a chapter on consciousness where free will is not mentioned. In a second stage, subjects were given a computer-based math test which featured an opportunity to cheat. The results seem to show that participants cheated more frequently after reading the anti-free-will excerpt than after reading the control one. Vohs and Schooler conclude that an exposure to deterministic messages gives place to a weakening of free-will beliefs, thus increasing the likelihood of unethical actions (2008, 53-54).

However, these results could hardly be used as evidence for Smilansky's catastrophism. Sommers (2010) has rightly noticed that Vohs and Schooler's study assumes that "our behavior just after hearing that a cherished belief is false has [some] bearing on how we would act after further reflection." But there is no study that records this correlation. Perhaps, then, this study "may merely have shown that people should not become hard determinists 15 min before they submit their tax return." Indeed, these are short-term implications of a limited denial of free will, whereas Smilansky's claim regards "the long-term implications of a widespread denial of free will" (Sommers 2010, 207). Moreover, Vohs and Schooler's results tell us very little about the effect that the weakening of free-will beliefs might have on more significant moral behaviors—"cheating on a math test is one thing; robbing banks is another," Nahmias rightly notices (forthcoming, note 16).

The problem is that, as Nadelhoffer and Feltz warn us, it is really very difficult to test Smilansky's claim. In fact, suppose you run an experiment in which you ask believers in freedom and responsibility to predict how they would behave if they came to think that their beliefs are the product of self-illusory mechanisms. The trouble is that:

> ... there are several reasons to suspect that simply asking people what they would do if they came to abandon some of their most fundamental beliefs would produce unreliable data. One of the primary shortcomings of this kind of study is that we have good reason to suspect that people will not be good judges of how they

> would behave if they no longer believed in [libertarian free will] and [ultimate moral responsibility] (Nadelhoffer & Feltz 2007, 210).

Numerous findings of social psychology invite us to distrust the reliability of predictions that people can make about their own future behavior. Nadelhoffer and Feltz (2007, 210-11) remind us that prior to the famous shock studies by Milgram (1963), Milgram polled his colleagues and Yale University senior-year psychology majors to establish what they thought would be the maximum shock administered by the participants in the experiment: their predictions ranged anywhere from 195 volts to 300 volts. In contrast, no subject stopped before 300 volts, and 26 of 40 participants administered what they believed to be the maximum shock of 450 volts. These results suggest that the degree of obedience among participants was far higher than people had predicted. Moreover, Nadelhoffer and Feltz notice, "[p]resumably, had people been asked to predict how *they themselves* would behave rather than being asked how *other people* would behave, their predictions would have been even less reliable" (note 11). In support of this claim they quote an experiment by Gilbert *et al.* (2002) that shows that in predicting we often overestimate the effect that negative events will have on our lives.

Alternatively, we could try to carry out a longitudinal research on the behavior of people who have denied free will and moral responsibility for long time. In this case, the trouble—Sommers says—is that "those individuals are few and far between" (2010, 207). On the other hand, we could get around the logistic impracticability of such a project by looking back to the aforementioned historical cases of the Lutheran and Calvinist free-will deniers. Moreover, "Smilansky himself is presumably perfectly able to live a morally acceptable and personally gratifying life, filled with meaningful choices and loving relationships" (Nadelhoffer & Feltz 2007, 211-2). And the same seems to hold for other skeptical philosophers like, say, Diderot and Spinoza. Why then should we be afraid that the same disillusionment will cause inauspicious and long-term consequences to the masses?

2. Cooperating and Punishing

According to a common opinion, our ideal of justice—and the punitive practices of the legal systems that purport to express it—presupposes the retributivist conception of punishment. Several authors (including Greene & Cohen 2004), however, disagree with this opinion, by arguing that the conception of punishment that better embodies our ideal of justice is the utilitarian one.

Greene and Cohen claim that we already have good reasons to think that free will is illusory, but (differently from Gazzaniga) they think that belief can have extremely beneficial and progressive consequences for society overall. This is because, according to them, it makes no sense to punish individuals

who could not act differently from how they have in fact acted, since they are genetically and neurophysiologically determined: "At this time, the law deals firmly but mercifully with individuals whose behaviour is obviously the product of forces that are ultimately beyond their control. Someday, the law may treat all convicted criminals this way. That is, humanely." (Greene & Cohen 2004, 1784).

Green and Cohen's view, however, is simplistic. Giving up the idea of free will implies that one has to abandon the galaxy of notions that essentially depends on the idea of freedom—such as responsibility, desert, merit, and guilt.[8] But, without these notions, one is left with a purely utilitarian theory of punishment, according to which any punishment could be inflicted to anybody, as long as the general utility were increased[9]. In this scenario, ideas such as punishment or blame are unjustified: there is no justice ever to be restored, no responsibility to be considered. The only things that have to be done are those which are useful: to rehabilitate the wrongdoer back into social life, set examples that can deter other potential wrongdoers, protect society against dangerous individuals. Nothing more than that.

This view does not fit well with our ideals of justice. But it should be noticed that, if put into effect, it would not generate the collapse of our judiciary system, contrary to what has been argued by several authors (Gazzaniga 2008; Caruana 2010). In fact, there are many cases of judicial systems that are based on similar kinds of views. Clear approximations to the ideal of a purely utilitarian judicial system are offered, for example, in some East-Asian countries, which proudly defend the idea of the so-called "Asian values" – that is, the idea that the prosperity of the community has always to have priority on what is good for the individuals. In this perspective, punishing an innocent person may sometimes be acceptable or even required: this may for example happen in case of dangerous social disorders that could be stopped by finding a "scapegoat" (the relevant consideration in these cases is whether the utility for society would be higher than the sufferance of the innocent person).[10]

Still, a utilitarian legal system that allows such easy ways out for social dilemmas seems intuitively objectionable. The point, however, is not that these kind of systems would generate anarchy, as Gazzaniga and Caruana claim (on the contrary, they offer good, if too easy, answers to the menace of anarchy), but that they do not respect the fundamental fairness requirements that our intuition of justice carries. It is in order to avoid such very unpalatable results that today even many of the most important utilitarians

[8] An exception to the classic view that moral responsibility requires free will is Frankfurt (1969), which is criticized in De Caro (2004), cap. IV.

[9] This is not the place to argue that the attempt of answering this kind of problem by the so-called "rule-consequentialism" is not satisfying. On that, cf. Habib (2008).

[10] The proposal of developing a "rule utilitarianism", instead of the more classic "act utilitarianism" is of not help here, since the same problem would raise again in a different form (see: Lyons 1965).

accept, very reasonably indeed, the idea of "negative retribution", according to which nobody can be punished who is not guilty. The classic advocate of this view is H. L. A. Hart (1968) who, while conceiving justification of punishment in purely utilitarian terms, regarded negative retribution as a "limiting principle", aimed at constraining the distribution of punishment by avoiding patent injustices. In this perspective, no victim-perpetrator should be punished if he or she is not morally responsible for a criminal action, even if the punishment would provide a benefit to society overall[11].

And this shows that moral responsibility is relevant not only for all retributivist views (which claim that moral responsibility is a necessary and sufficient condition for punishment), but also for the utilitarian views of the Hartian family (since they claim that moral responsibility, if not a sufficient, is at least a necessary condition of punishment). The crucial point however, is that according to most authors, in order to attribute moral responsibility to people, we have to attribute to them free will as well (either in the compatibilist or in the libertarian form). But Gazzaniga and Caruana, on the pessimistic side, and Greene and Cohen, on the optimistic side, defend an illusionist view of free will; in their views, therefore, all concepts that are connected with free will (such as merit, responsibility, guilt) should be abandoned. In this way, none of these authors can appeal to the notion of negative retribution, which requires the concept of moral responsibility (or some other concepts of the free will galaxy). It follows that those views—by using purely utilitarian parameters for both the justification and distribution of punishment—imply a conception of punishment that, if not as apocalyptic as Gazzaniga and Caruana think, are certainly very far from being as progressive as Greene and Cohen maintain.

Greene and Cohen's proposal of an utilitarian re-engineering of the legal system can be opposed also by referring to some experimental findings that concern our *intrinsic* motivation to punish. There is robust anthropological and sociological evidence that shows that the members of a community react to norm violations with both punitive emotions (e.g., anger, contempt, and disgust)[12] and punitive behaviors (e.g., criticism, condemnation, avoiding, exclusion, or aggression). These informal ways of punishing norm violations seem to be cultural universals. Nevertheless, it might be objected that in these cases there is no intrinsic motivation to punish. That is, it may be that one wants to punish someone who violated a norm for some egotistical instrumental reasons—e.g., to warn the transgressors as a deterrent from committing the violation again. Against this claim, we can refer to two studies

[11] It should be noted that Hart is a utilitarian, even if he accepts the idea of negative retribution, since he refuses the idea of positive retribution—i.e., he does not claim that all people who deserve punishment should be punished, whatever consequences their punishment may have.

[12] For a review, see: Haidt (2003). These punitive emotions are connected to the aforementioned Strawson's "reactive attitudes".

– one in social psychology, and the other in experimental economy—that serve as a source of evidence for the hypothesis that in some cases the motivation to punish is genuinely intrinsic[13].

In a study by Haidt and Sabini (2000) participants were shown with clips from Hollywood movies where a norm violation occurred. In a second stage subjects were asked to rate several alternative endings. The result was that subjects were likely to give higher ratings of the endings in which the perpetrators (i) were made to suffer, (ii) knew that their suffering was fair repayment for the violation, (iii) suffered as much as the victim, and (iv) their suffering involved a public humiliation. Most importantly, among the alternative endings there was one in which the perpetrators realized they did wrong and felt a genuine remorse that put them on a path of redemption and rehabilitation. That subjects were unsatisfied by this ending suggests that their motivation to punish cannot be characterized in terms of selfish instrumental ends (e.g., avoiding of being harmed by the perpetrator in the future); rather, their motivation to punish was genuinely intrinsic.

Moreover experimental economy offers systematic findings relevant to the hypothesis of a natural inclination to retributive punishment. The Ultimatum Game is a simple bargaining situation in which the experimenter provides a pair of anonymous subjects with a sum of real money (e.g., $100) for a one-shot interaction (Guth, Schmittberger, & Schwarze 1982). One of the pair (A) has to offer a portion of the sum to a second player (B): A can give B from $1 to $99, as she likes. If B accepts the offer, the sum is split as proposed; if B rejects it both players receive nothing. According to classical game theory, since A takes B to be a rational agent for whom any amount of money has a positive utility, A anticipates that B will accept any offer >0. A should therefore offer the smallest possible amount, in order to keep as much money as possible, and B should accept any proposed amount, because "few is better than nothing". However, this is not what happens. Although the specifics vary across culture and setting, the typical results are that A makes offers of 40 to 50% and B rejects offers <20%. These findings suggest that B is sensitive to unfairness and punishes A's inequitable offers, although punishment may be costly for B and yield no material gain. We find here a natural desire to pay a cost in order to send the signal "you had to be more cooperative". In other words, we have here an inclination to "altruistic punitiveness" (Fehr & Gächter 2002; Boyd *et al.* 2003).

Fehr and Gächter offer cogent evidence for the hypothesis that "cooperation flourishes if altruistic punishment is possible, and breaks down if it is ruled out" (2002, 137). In a "public goods" experiment participants had the option to co-operate by contributing significantly to a common fund or defect by not contributing (Fehr & Fischbacher 2004). During the first ten trials, no punishment was allowed. During trials 11-20, group members could punish each other after they observed each member's contribution level. At

[13] We are following Sripada (2005), Stich & Sripada (2006), Nichols (2007, 2008) here.

the beginning of the experiment co-operation rates of roughly 50% of the endowment (=20 monetary units) were observed, but the level of co-operation decreased over time. The majority of subjects contributed nothing to the public good in the trial 10, and the rest contributed little. In period 11, the subjects were informed that a new experiment would start in which they would have the opportunity to punish the other group members (but with a cost for themselves)[14]. This modification immediately increased co-operation levels to 65% of the endowment. Then, over time, cooperation rose dramatically, and eventually almost 100% cooperation was attained.

Fehr and Gächter's study aims to show that "cooperation flourishes if altruistic punishment is possible, and breaks down if it is ruled out." Therefore, if this line of research takes root, the project of reforming jurisprudence without retributivist punishment will seem "a dangerous cause" (Nichols 2008). And this is still more important since, as Hart has clearly pointed out, the retributivist intuition can act as a limit to the arbitrary applications in the distribution of punishment.

References

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. 2009. "Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness." *Personality and Social Psychology Bulletin* 35(2): 260-268.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. 2003. "The Evolution of Altruistic Punishment." *Proc. Natl. Acad. Sci. USA* 100: 3531-3535.

Caruana, F. 2010. "Due problemi sull'utilizzo delle neuroscienze in giurisprudenza." *Sistemi Intelligenti* 2: 337-346.

Dennett, D. C. 2003. *Freedom Evolves.* London: Penguin Books.

Fehr, E. & Gächter, S. 2002. "Altruistic Punishment in Humans." *Nature* 415: 137-140.

Fehr, E. & Fischbacher, U. 2004. "Social Norms and Human Cooperation." *Trends in Cognitive Sciences* 8: 185-190.

Flanagan, O. 2002. *The Problem of the Soul*. New York: Basic Books.

Gazzaniga, M. 2008. "The Law and Neuroscience." *Neuron* 60: 412-415.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. 2002. "Durability Bias in Affective Forecasting." In Gilovich, T., Griffin, D., & Kahneman, D. (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge*:* Cambridge University Press.

Greene, J. & Cohen, J. 2004. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical Transactions of the Royal Society of London,* B 359: 1775-1785.

[14] Every monetary unit invested into punishment decreased the punished member's monetary payoff by 2-4 monetary units.

Guth, W., Schmittberger, R., & Schwarze, B. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization* 3: 367-88.

Haidt, J. 2003. *The Moral Emotions.* In Davidson, R. J., Scherer, K. R., & Goldsmith, H. H. (Eds.), *Handbook of affective sciences*. Oxford: Oxford University Press.

Haidt, J. & Sabini, J. 2000. *What Exactly Makes Revenge Sweet?*. Unpublished manuscript.

Hart, H. L. A. 1968. *Punishment and Responsibility.* Oxford: Oxford University Press.

Kane, R. 1999. "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism." *Journal of Philosophy* 96: 217-240.

Kim, J. 1998. *Mind in a Physical World.* Cambridge (Mass.): MIT Press.

Knobe, J. & Nichols, S. 2008. *Experimental Philosophy.* Oxford*:* Oxford University Press.

Michaels, A. C. 2004. "Fastow and Arthur Andersen: Some Reflections on Corporate Criminality, Victim Status, and Retribution." *Ohio State Journal of Criminal Law* 2: 551-71.

Milgram, S. 1963. "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology* 67: 371-378.

Nadelhoffer, T. & Feltz, A. 2007. "Folk Intuitions, Slippery Slopes, and Necessary Fictions: An Essay on Saul Smilansky's Illusionism." *Midwest Studies in Philosophy* 31: 202-213.

Nahmias, E. 2006. "Folk Fears About Freedom and Responsibility: Determinism vs. Reductionism." *Journal of Cognition and Culture* 6(1-2): 215-38.

———. (forthcoming, in press). "The Psychology of Free Will." In Prinz, J. (Ed.), *The Oxford Handbook on Philosophy of Psychology*. Oxford: Oxford University Press. Retrieved from: <http://www2.gsu.edu/~phlean/papers.html>.

Nahmias, E. & Murray, D. (forthcoming, in press). "Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions." In Aguilar, J., Buckareff, A., & FrankisH, K. (Eds.), *New Waves in Philosophy of Action*. Palgrave-Macmillan, <http://www2.gsu.edu/~phlean/papers.html>.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. 2005. "Surveying Free Will: Folk Intuitions About Free Will and Moral Responsibility." *Philosophical Psychology* 18(5): 561-584.

———. 2006. "Is Incompatibilism Intuitive?." *Philosophy and Phenomenological Research* 73(1): 28-53.

Nichols, S. 2006. "Folk Intuitions About Free Will." *Journal of Cognition and Culture* 6: 57-86.

———. 2007. "After Incompatibilism: A Naturalistic Defense of the Reactive Attitudes." *Philosophical Perspectives* 21: 405-428.

———. 2008. “How Can Psychology Contribute to the Free Will Debate?” In Baer, J., Kaufman, J., & Baumeister, R. (Eds.), *Are We Free?* Oxford: Oxford University Press.

Nichols, S. & Knobe, J. 2007. “Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions.” *Nous* 41(4): 663-685.

Roskies, A. & Nichols, S. 2008. “Bringing Moral Responsibility Down to Earth.” *Journal of Philosophy* 105, 7: 371-388.

Russell, P. 1995. *Freedom and Moral Sentiment.* Oxford*: Oxford University Press.*

Smilansky, S. 2000. *Free Will and Illusion*. Oxford*:* Oxford University Press.

———. 2002. "Free Will, Fundamental Dualism, and the Centrality of Illusion." In Kane, R. (Ed.), *Oxford Handbook of Free Will*. Oxford*:* Oxford University Press.

Sommers, T. 2010. “Experimental Philosophy and Free Will.” *Philosophy Compass* 5(2): 199-212.

Sripada, C. 2005. “Punishment and the Strategic Structure of Moral Systems*.”* *Biology and Philosophy* 20: 767-89.

Sripada, C. & Stich, S. 2006. “A Framework for the Psychology of Norms.” In Carruthers, P., Laurence, S., & Stich, S. (Eds.), *The Innate Mind: Culture and Cognition.* Oxford: Oxford University Press.

Strawson, G. 1986. *Freedom and Belief.* Oxford: Oxford University Press.

Strawson, P. F. 1962. “Freedom and Resentment.” *Proceedings of the British Academy* 48: 1-25.

Turner, J. & Nahmias, E. 2006. “Are the Folk Agent-Causationists?.” *Mind and Language* 21(5): 597-609.

Vohs, K. D. & Schooler, J. W. 2008. “The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating.” *Psychological Science* 19: 49-54.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. 2006. “Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility.” *Cognition* 100: 283-301.

Mario De Caro (Roma)
Massimo Marraffa (Roma)

Free Will and Retribution Today

Abstract: The paper addresses two issues that have been recently debated in the literature on free will, moral responsibility, and the theory of punishment. The first issue concerns the descriptive project, the second both the substantive and the prescriptive project. On theoretical, historical and empirical grounds, we claim that there is no rationale for fearing that the spread of neurocognitive findings will undermine the ordinary practice of responsibility attributions. We hypothetically advocate two opposite views: (i) that such findings would cause the collapse of all punitive practices; (ii) that, on the contrary, such findings would open the way to more humane forms of punishment, which would be justified on purely utilitarian grounds. We argue that these views are both wrong, since whereas a sound punitive system can be justified without any reference to moral responsibility, it will certainly not improve the humaneness of punishment.

Keywords: Free will, moral intentions, responsibility problem, punitive practices revised, neurocognitive findings

Doi: 10.14746/eip.2014.2.2