Ethic Theory Moral Prac (2009) 12:511–524
DOI 10.1007/s10677-008-9143-5

Responsibility and the Brain Sciences

Felipe De Brigard · Eric Mandelbaum · David Ripley

Accepted: 20 November 2008 / Published online: 24 December 2008

Abstract Some theorists think that the more we get to know about the neural underpinnings of our behaviors, the less likely we will be to hold people responsible for their actions. This intuition has driven some to suspect that as neuroscience gains insight into the neurological causes of our actions, people will cease to view others as morally responsible for their actions, thus creating a troubling quandary for our legal system. This paper provides empirical evidence against such intuitions. Particularly, our studies of folk intuitions suggest that (1) when the causes of an action are described in neurological terms, they are not found to be any more exculpatory than when described in psychological terms, and (2) agents are not held fully responsible even for actions that are fully neurologically caused.

Keywords Responsibility · Neuroscience · Free will · Experimental philosophy · Mental illness · Law

1 Introduction

If you find yourself living in a small apartment with your healthy parents when you are 36, presumably something has not gone according to plan. Take Curtis, a hypothetical 36-year-old unemployed college graduate who currently lives with his parents. Imagine one day his father walks into his room after Curtis has just come home from a visit to the therapist, and that he finds Curtis in a familiar scene: in the dark, lying awake on his bed, staring at the ceiling. The father asks Curtis if he is OK, and if he plans on going out and looking for a job tomorrow. Curtis replies that he has just received bad news from his doctor. He found out that he is depressed; that's why he hasn't gotten out of bed and looked for a job in the past few months. Without either party knowing the etiology of the depression, it is not hard to imagine this

All authors contributed equally and are listed alphabetically.

F. De Brigard · E. Mandelbaum (✉) · D. Ripley
Department of Philosophy, UNC/Chapel Hill, CB #3125, Caldwell Hall, Chapel Hill, NC 27599-3125, USA
e-mail: ericman@email.unc.edu

situation as one in which the father thinks that Curtis is lazy and is responsible for his current unemployed predicament.

But imagine the details were slightly different: imagine that Curtis went to a neurologist instead of a therapist, and that the neurologist discovered that Curtis had an aneurysm. The neurologist informed Curtis that the aneurysm has caused the depression and the ensuing lethargy that has, in turn, caused Curtis to lie awake on his bed effectively too paralyzed to look for a job. In this case it's not hard to imagine Curtis's father expressing sympathy for Curtis's plight, instead of showing "tough love." Intuitively, it seems that fathers are more understanding towards their unemployed middle-aged sons when the reason for their unemployment has an identifiable neurological cause. Contrarily, it seems that fathers' sympathies dwindle when that cause is identified as a somewhat nebulous psychological state, like depression.

One might think that this apparently ubiquitous reasoning is caused by the way people think about neurological and psychological states in general. Perhaps people think that we are more responsible for our psychological states than we are for our neurological states. This hypothesis receives some support from a study reported in Nahmias (2006). In exploring the folk's intuitions concerning free will and responsibility, Nahmias hypothesized that the main variable controlling participants' intuitions had to do with a fear of what he calls "bypassing."[1] Nahmias supposed that if an agent's conscious choices are bypassed, then we should expect participants to judge that the agent in question is not responsible for his or her actions. To operationalize the notion of "bypassing," Nahmias presented participants with one of two scenarios. Both scenarios take place in a deterministic universe, but in one scenario, agents' actions are completely caused by their psychological states (which are in turn completely caused themselves), and in the other, agents' actions are completely caused by their neurological states (which are in turn completely caused themselves). When the agents' actions were determined by their psychological states, participants judged that the agents had free will and deserved praise or blame for their actions, but when the actions were determined by agents' neurological states, participants judged that the agents did not act of their own free will and did not deserve praise or blame for their actions. These results are quite surprising; it seems that it is not the deterministic universe per se which is dictating the participants' intuitions, but rather some type of difference in how people reason about psychological and neurological states (and perhaps their relation to bypassing, as Nahmias hypothesizes).

This result sets up a starker worry: if people do reason differently about neurological and psychological states, then what will happen when our neuroscience advances to the point that all explanation of behavior can be couched in neurological terms?[2] This worry is forcefully presented in Greene and Cohen (2004), where they speculate that our common-sense ideas of free will will be imperiled as neuroscience advances. They argue that once we are brought up in a world where neurological explanations of behavior are commonplace, the "my brain made me do it" (Gazzaniga 2005) defense will be enough to cause juries to judge that defendants are not morally responsible for their actions. They also argue that societal judgments of moral responsibility underpin the laws governing legal responsibility. Consequently, they suggest that the advance of neuroscience may force the

[1] Nahmias writes, "Most people do not regard psychological determinism to be a threat to free and responsible action but most people do regard reductionistic pictures that suggest `bypassing' to be a threat to free and responsible action" (p. 233–234).

[2] Assuming, of course, that this situation can arise. There are some who suppose that even when we have a completed neuroscience the explanations of most types of behavior will still not be couched in neurological terms (e.g. Fodor 1974).

law to make some wide-ranging changes, since society will not tolerate laws that seem out of step with their moral judgments. In particular, they worry that juries will be willing to acquit defendants willy-nilly for the defendants' actions regardless of what the law requires. We won't take issue here with the part of Greene and Cohen's argument connecting judgments of moral responsibility to the law; our results bear more directly on judgments of moral responsibility themselves.

With this background in hand, we set out to take a first step at finding out how the folk make judgments of moral responsibility when an agent's behaviors are caused by either a neurological or a psychological illness. This question is interesting and important in itself, for many people are forced to make these types of judgments in settings where the consequences are dire for the defendants. Yet this question might also lend a bit of insight into the free will and determinism debate. If the ways people reason about psychological and neurological states differ, then presumably they should also differ when thinking about psychological and neurological illnesses. Accordingly, one may tentatively want to predict that we will see a difference in the way people reason about psychological and neurological illnesses. Moreover, if an agent's actions are outside of her control because of a psychological or neurological illness, then prima facie we have a case of determined action and the folk's judgments in these cases should suggest how the folk reason about responsibility, free will, and determinism. Lastly, seeing how the folk reason about agents whose behavioral etiology is described in neurological terms should tell us a bit about Greene and Cohen's hypothesis. If the folk still hold agents responsible for their actions even when their actions are determined by their neurological states, then this would give us some reason for being suspicious of Greene and Cohen's hypothesis and thus suspicious of the thought that the advance of neuroscience will necessitate a change in our legal system.

2 Study 1

In our first experiment we set out to see whether people would hold someone with a neurological illness less responsible than someone who had a psychological illness. Participants were randomly assigned to one of six conditions, with conditions grouped to form three pairs (conditions 1, 2 and 3 each had a psychological and a neurological variant). The vignettes read as follows:

Psychological Condition 1

Fred, a middle aged man, constantly finds himself thinking about pre-pubescent boys in sexual ways. Fred doesn't want to have these thoughts, but these are constant events in his daily life. From Fred's apartment he can peek into the bathroom in an apartment next to his. One day Fred sees a pre-pubescent boy showering and watches the boy until he finishes his shower. Fred has been diagnosed with **Huebner's delirium, a psychological condition characterized by excessive psychomachinations.**
On the following scale, how responsible is Fred for watching the boy shower?

Neurological Condition 1

Fred, a middle aged man, constantly finds himself thinking about pre-pubescent boys in sexual ways. Fred doesn't want to have these thoughts, but these are constant events in his daily life. From Fred's apartment he can peek into the bathroom in an apartment next to his. One day Fred sees a pre-pubescent boy showering and watches the boy until he finishes his shower. Fred has been diagnosed with **Cilifibrial Hermatosomes, a neurological condition characterized by dendritic hepatocytes.**
On the following scale, how responsible is Fred for watching the boy shower?

Psychological Condition 2

Cliff is 45 and has never had problems with his eyes. A few weeks ago Cliff's family started to realize that Cliff was becoming clumsier. At his family's behest Cliff went to see the doctor. The doctor proceeded to tell Cliff that he is blind. Cliff does not agree with his doctor, frequently protesting that he can in fact see and that he would know better than anyone else whether he was or was not blind. One day Cliff decided to drive to the supermarket and when pulling out of his driveway he crashed into a passing cyclist, sending him directly to the hospital. Cliff has been diagnosed with Anosognosia**, a psychological syndrome characterized by excessive psycholocutions**. He honestly denies having any impairment.
On the following scale, how responsible is Cliff for crashing into the passing cyclist?

Neurological Condition 2

Cliff is 45 and has never had problems with his eyes. A few weeks ago Cliff's family started to realize that Cliff was becoming clumsier. At his family's behest Cliff went to see the doctor. The doctor proceeded to tell Cliff that he is blind. Cliff does not agree with his doctor, frequently protesting that he can in fact see and that he would know better than anyone else whether he was or was not blind. One day Cliff decided to drive to the supermarket and when pulling out of his driveway he crashed into a passing cyclist, sending him directly to the hospital. Cliff has been diagnosed with Anosognosia**, a neurological disorder characterized by having damage to the occipital lobe of one's brain**. He honestly denies having any impairment.
On the following scale, how responsible is Cliff for crashing into the passing cyclist?

Psychological Condition 3

Jennifer, 41, separated from her husband a year ago. After a very painful and difficult divorce her lawyer suggested that she stay away from her ex-husband as much as possible. And so she did. A week ago, however, they ran into each other in an open parking lot in the middle of the day. Nobody was around. When she saw him approaching, she pulled her pepper-spray gas out of her purse and proceeded to spray it into her ex-husband's eyes. Due to major eye irritation, her ex-husband had to go to the emergency room, where he decided to sue his ex-wife on the charge of personal injury. She claims to have not recognized her husband, mistaking him for a mugger. Recently, Jennifer was diagnosed with Prosoponomia, a **psychological** condition where one may fail to recognize faces.
On the following scale, how responsible is Jennifer for having sprayed the gas at her ex-husband?

Neurological Condition 3

Jennifer, 41, separated from her husband a year ago. After a very painful and difficult divorce her lawyer suggested that she stay away from her ex-husband as much as possible. And so she did. A week ago, however, they ran into each other in an open parking lot in the middle of the day. Nobody was around. When she saw him approaching, she pulled her pepper-spray gas out of her purse and proceeded to spray it into her ex-husband's eyes. Due to major eye irritation, her ex-husband had to go to the emergency room, where he decided to sue his ex-wife on the charge of personal injury. She claims to have not recognized her husband, mistaking him for a mugger. Recently, Jennifer was diagnosed with Prosoponomia, a **neurological** condition where one may fail to recognize faces.
On the following scale, how responsible is Jennifer for having sprayed the gas at her ex-husband?

Psychological Condition 1 and *Neurological Condition 1* are identical except for that in one condition Fred has been diagnosed with a (made up) psychological syndrome, and in the other condition Fred has been diagnosed with a (made up) neurological syndrome. The only other differences in this pair are a) the name of the fake syndrome and b) the name of the fake symptoms. *Psychological Condition 2* and *Neurological Condition 2* are also identical except for a) whether the named syndrome was characterized as a neurological syndrome or a psychological syndrome and b) the symptoms of the syndrome. *Psychological Condition 3* and *Neurological Condition 3* were also near-identical, the only difference between the two prompts being the exchange of the word 'psychological' for the word 'neurological.' Titles and boldface font did not appear in the actual vignettes (the boldface has been inserted to highlight the differences between the pairs).

Participants were 217 undergraduate students in introductory philosophy classes at the University of North Carolina at Chapel Hill. The surveys were given out to the participants at the end of their class and their participation was voluntary. Each participant received only one vignette and was given a scale that ran from 1 to 7, with 1 corresponding to 'less responsible' and 7 corresponding to 'more responsible.'

The means for each vignette are given below in Fig. 1.[3] Four main points are worth mentioning. First, and most importantly, there was no overall statistically significant difference between the neurological and psychological conditions. Second, there were no statistically significant differences within the pairs. Third, and quite strikingly, although the effect was not statistically significant, on average participants were more apt to assign responsibility in the neurological condition than in the psychological condition for each vignette. This clearly casts doubt upon the intuition that neurological descriptions of behaviors motivate more exculpatory judgments than psychological descriptions. Finally, overall, participants did not find that just having a psychological or neurological syndrome was a large mitigating factor; with answers well above the mid-point, participants still found the characters to be more or less responsible.

However, note that in this study we didn't explicitly state that the syndrome that the character suffered from was actually the cause of the behavior in question. To see whether this was a factor that may have been influencing our participants' reasoning we ran a follow-up study.

3 Study 2

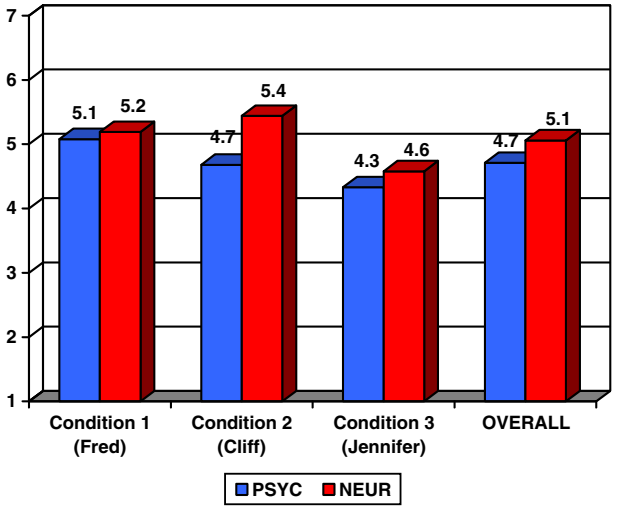
Study 2 was run exactly like Study 1 except that in Study 2 a) we make it clear that the character in the vignette couldn't have known about his syndrome before the action took place and b) we changed the wording of the scale of responsibility—it ran from 'not responsible' (1) to 'very responsible' (7). The titles and boldface fonts were not included in the original vignettes (boldface is once again added to highlight the differences between the prompts). The vignettes are reproduced below.

Psychological Condition

Dennis and John have been friends for thirty years who always meet for a weekly walk. Dennis has been away on vacation for a month and so the friends have not been able to go on their walk until last week. On their walk last week they passed a jogger on their normal trail. Seemingly unprovoked. Dennis ran up to the jogger and punched him in the stomach multiple times. Shortly after this incident Dennis was diagnosed with a **psychological** illness that causes him to manifest uncontrollably aggressive behavior which in turn caused him to hit the jogger. On a scale of 1–7 how responsible is Dennis for hitting the jogger?

[3] The data was subjected to a 3 (scenario: Jennifer vs. Cliff vs. Fred) ×2 (condition: psychological vs. neurological) between participants ANOVA. There was no significant difference between the neurological scenarios and the psychological scenarios, F (1, 211)=3.1, $p>.05$. We did find a significant difference between the characters, F (2, 211)=4.0, $p<.05$. The significant difference was driven by the Jennifer vignette; participants rated her as less responsible than the other characters, regardless of whether her illness was of a psychological or neurological nature. There was no significant interaction effect between the conditions and scenarios.

Fig. 1 Study 1



Neurological Condition

Dennis and John have been friends for thirty years who always meet for a weekly walk. Dennis has been away on vacation for a month and so the friends have not been able to go on their walk until last week. On their walk last week they passed a jogger on their normal trail. Seemingly unprovoked, Dennis ran up to the jogger and punched him in the stomach multiple times. Shortly after this incident Dennis was diagnosed with a **neurological** illness that causes him to manifest uncontrollably aggressive behavior, which in turn caused him to hit the jogger. On a scale of 1–7, how responsible is Dennis for hitting the jogger?

The vignettes were given to 60 UNC-Chapel Hill undergraduates at the end of their philosophy class. The means for the two conditions were exactly the same: they both received a mean score of 3.8 (see Fig. 2).[4] Once again, there was no significant difference between the *Neurological* and *Psychological Conditions*.[5] Moreover, participants still answered, on average, that the agent was 'kind of responsible.'

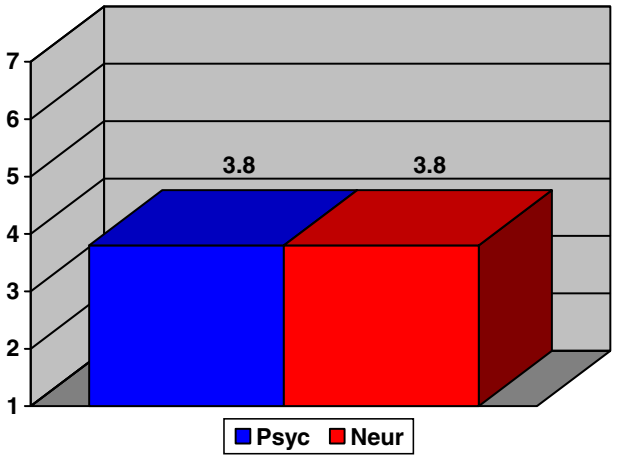
Some may find it startling that although our participants found the syndromes to be somewhat mitigating factors, they did not find them completely exculpatory. The vignettes explicitly mention that his syndrome caused Dennis to perform the action and clearly imply that he had no knowledge of having the syndrome at the time of the action, yet our participants still found him to be responsible. Prima facie, these results are puzzling. They suggest that our participants believe that people are still responsible even if a standing neurological or psychological condition causes their action.

In sum, the results of our studies give support to the following claim: people do not judge a behavior that is depicted as having been caused by a neurological condition differently than a behavior that is depicted as having been caused by a psychological

[4] These middling scores may be taken to reflect the uncertainty that the participants had about the case. However, we did allow our participants to comment on the vignettes if they cared to and we did not receive any feedback which made us think that they were particularly confused. However, this possibility cannot be ruled out. We thank an anonymous reviewer for pointing this possibility out to us.

[5] $t(58)=.05$; $p>0.9$

Fig. 2 Study 2



condition. This is not what we might have at first expected, especially given the theoretical background mentioned above, provided by Nahmias, Gazzaniga, and Greene and Cohen.

4 Study 3

There remain a few concerns about these studies, however. First, the studies were run on students in introductory philosophy classes, a population which may not be indicative of the population at large; one might also worry that these students may have been contaminated by discussing these issues in class. These concerns may not be so dire: we ensured that our participants were in introductory classes that had not previously covered issues pertaining to free will, responsibility, or the law, and so we are fairly confident that our population pool was not contaminated. However, we are sympathetic to the notion that students who take introductory philosophy class are not indicative of people at large.

Second, it's important to the interpretation we've offered that participants in the psychological conditions are really thinking of the character in their vignette in psychological terms, and that participants in the neurological conditions are really thinking of the character in their vignette in neurological terms. But the vignettes used in the first two studies might not be enough to guarantee this. After all, there were relatively minimal differences between the conditions in these studies.

In particular, it might be that the presence of the single word "psychological" or "neurological" isn't enough to get participants to really understand the case in either psychological or neurological terms; but that was the only difference between the conditions in our Study 2 and the third pair in our Study 1. The first two pairs in Study 1 offered a bit more to make the vignettes clear, but it would be hard to make ironclad the claim that this was sufficient. "Excessive psychomachinations" (e.g.) sounds psychological to *us*, and "dendritic hepatocytes" neurological, but how do these phrases sound to our participants? It's hard to say with certainty.

Third, it's possible, for all we've seen, that participants judge the characters with neurological and psychological illnesses the same, not because of any parallel between neurology and psychology, but instead simply because they have illnesses. Maybe illnesses are judged differently from other ways in which behavior might be caused. More generally,

perhaps the abnormality of the causal pathways mattered more than their psychological or neurological nature.

To resolve these issues, we conducted a third study. The core idea is similar to that of our first two studies, but the vignettes were given to a wider variety of students, they were much more explicit in using psychological or neurological language to explain a behavior's causal chain, and the explanations didn't involve illnesses or abnormalities at all. Once again, participants were asked to rate the character's moral responsibility on a scale from 1 to 7. The names of the conditions and the boldface text were not presented on the original prompts.

Psychological Condition

Jerry is usually an attentive driver. Last Tuesday, however, he failed to stop at a red light and ran into another car, injuring the driver. He was distracted; that morning the documents for his divorce finally came through and he signed his resignation letter. Many memories clouded his mind and he felt pretty sad. His sadness led to anxiety, which **in turn caused him to experience increasing feelings of apprehension, concern, and nervousness. As a result, he was absentminded and had difficulty focusing on visual information. This, in turn, made him pay less careful attention to the visual information coming from the road, which is necessary for conscious awareness of visual events.** That is why Jerry failed to notice the red light.
On a scale of 1–7, with 1 being 'NOT RESPONSIBLE' and 7 being 'VERY RESPONSIBLE', how responsible is Jerry for the car accident?

Neurological Condition

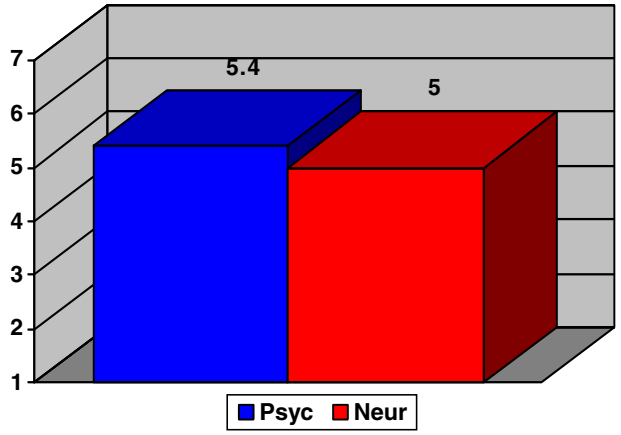
Jerry is usually an attentive driver. Last Tuesday, however, he failed to stop at a red light and ran into another car, injuring the driver. He was distracted; that morning the documents for his divorce finally came through and he signed his resignation letter. Many memories clouded his mind and he felt pretty sad. His sadness led to anxiety, which **caused a sudden reduction of his levels of dopamine (a neurotransmitter whose function is to regulate blood pressure), and this in turn caused less oxygen to reach his occipital lobe, the part of the brain that processes visual information, making it operate more slowly and less efficiently than usual. This, in turn, caused less visual information from the road to pass into working memory, which is necessary for conscious awareness of visual events**. That is why Jerry failed to notice the red light.
On a scale of 1-7, with 1 being 'NOT RESPONSIBLE' and 7 being 'VERY RESPONSIBLE', how responsible is Jerry for the car accident?

Participants in this study were 66 students from the departments of theology and communication in the Universidad Javeriana in Bogotá, Colombia, and 53 business students in New York City (these were students at Baruch College and non-university GMAT students at a test preparation company).[6] Of these 119 participants, 58 were randomly assigned to the Psychological Condition, and 61 to the Neurological Condition. Yet again, there was no significant difference in the mean responses (the averaged means are reproduced below in Fig. 3).[7] Note that in this study we make it quite explicit that the causal chain is either psychological in nature or neurological in nature. However, even the

[6] The Colombian students were students from varied majors in a critical writing class, while the New York students were business students who were either in a GMAT prep class or an introduction to philosophy class.

[7] We used a 2 (neurological vs. psychological) × 3 (Colombia, GMAT students, Baruch students) ANOVA. There was no significant difference between the neurological and the psychological condition, $F(5, 113)= 1.05$, $p=0.31$.There was a significant effect between the groups $F(5, 113)=3.16$, $p=0.046$, with the Colombian students holding the characters a bit less responsible regardless of condition. There was no significant interaction effect $F(5, 113)=0.015$, $p=0.985$

Fig. 3 Study 3



explicit stating of the neurological and psychological causal chains did not cause the participants to judge the two cases differently. Moreover, since this effect seems to be stable not only on students other than philosophy students, but cross-culturally, we suspect that people just do not find neurological etiologies to be more mitigating than psychological etiologies.[8]

5 Discussion

In sum, then, we found consistent results across all three studies: participants' responsibility judgments are not significantly affected by describing an action's causes in neurological as opposed to psychological terms. This result seems to hold whether or not the action in question is caused by an illness. Given our opening considerations, these results are quite surprising.

It doesn't seem that Nahmias would have predicted these results. After all, his study (mentioned above) found a striking asymmetry in folk judgments between neurologically- and psychologically-determined agents, while our study found no such asymmetry. How could he respond to the present data? Nahmias attributes the asymmetry in his cases to bypassing. "Bypassing," for Nahmias, is something of a term of art. The rough idea is that, in order for an agent not to be bypassed, that agent's behavior must be the causal result of some aspect of that agent's conscious mental life (or of the agent—Nahmias equates the agent to her conscious mental life (p. 217), but it's not clear whether he means to assert this equation himself or just put it in the mouth of the folk.)

Nahmias's hypothesis comes in two parts: (1) that when the causes of an agent's actions are described neurologically, participants will take the agent to be bypassed, and when the causes are described psychologically, participants won't take the agent to be bypassed; and (2) that participants judge bypassed agents to not be responsible, and to not have acted of their own free will. In our studies, though, there was no significant difference between judgments of responsibility in the psychological and neurological conditions. It seems that

[8] One could, of course, go further and claim that people actually don't think about neurological states differently than they think about psychological states. However, although we think that this is a live possibility, we leave the question open to further investigation.

at least part of Nahmias's hypothesis must go. One might try to retain part (2) of Nahmias's hypothesis by rejecting part (1); that is, one might reject the straightforward connection between the neurological/psychological distinction and the bypassed/not bypassed distinction. This might be a start, but we don't think it would be enough on its own.

Consider our first two studies, in which all the characters being judged were presented to participants as either psychologically or neurologically ill. It's not too hard to imagine that in these cases participants simply judged *all* the agents to be bypassed, in virtue of the agents' illnesses. If psychological and neurological illnesses equally count as bypassing an agent, then we should not expect a difference between them with regard to judgments of responsibility. This would explain the lack of asymmetry in our results, and it allows us to retain part 2) of Nahmias's hypothesis. Call this the "illness-is-bypassing" hypothesis, and suppose for the moment it's true. There's still a bit more to say.

First, in our initial three cases, all the agents were moved to act by their conscious mental states. Fred was moved by obsessive thoughts about young boys, Cliff by his disbelief in his own blindness, and Jennifer by her belief that her ex-husband was a mugger. None of them had their conscious mental states bypassed. On the illness-is-bypassing hypothesis, then, participants must have thought that even if the agents' conscious mental states weren't bypassed, the agents themselves were; the causally relevant mental states, they must have thought, were in some sense the illnesses' and not the agents' own. If this is correct, it suggests that Nahmias is too quick to equate the agent to the agent's conscious mental states, at least if he means to speak for the folk. If the illness-is-bypassing hypothesis is true, Fred, Cliff, and Jennifer were judged to have been bypassed, but their conscious mental states were still crucially involved. This suggests that Nahmias's picture of bypassing is incomplete.

Second, note that the illness-is-bypassing hypothesis is not enough to explain the results obtained in our third study, since Jerry is not presented there as pathological in any way, or even as abnormal in any way. To retain even part 2) of Nahmias's hypothesis, one would want some reason to think that Jerry is seen as equally bypassed in both conditions; we don't believe such a reason will be forthcoming.

But even if that issue were to be addressed, these studies put further pressure on Nahmias's view: if participants really did judge the agents to be bypassed, we should have expected responsibility judgments to be lower than they in fact were. After all, if an agent is bypassed, that agent didn't contribute causally to what occurred; they are no different from a bystander. So why did our participants hold them somewhat responsible? One hypothesis might be that the folk don't see psychologically or neurologically ill agents as completely bypassed. Perhaps bypassing comes in degrees, and illnesses of various sorts can produce partial bypassing.[9] Although this is a plausible thing to think, we do not think it is the sole reason for the relatively high responsibility judgments we obtained. Rather, we think these judgments are at least in part a result of our cases' specific and affect-loaded nature.

In studying folk intuitions about moral responsibility in deterministic universes, Nichols and Knobe (2007) found an interesting phenomenon. When presented with a fully deterministic universe, and asked whether it's possible for an agent in that universe to be "fully morally responsible for their actions", only 14% of participants answered yes. On the other hand, when given an identical description of the deterministic universe, and told about an agent in that universe, Bill, who kills his wife and family to be with his secretary, 72%

[9] Note, again, that this wouldn't explain the results of our third study.

thought that Bill was fully morally responsible! They conclude (and we concur) that the specificity of a case, and the affect aroused by it, can push the folk to judge agents more responsible than they would have otherwise. We hypothesize that this same phenomenon is what explains the relatively high responsibility judgments our participants gave.

To test this hypothesis, we ran a follow-up study which had two conditions (this study and its implications are reported and further discussed in Mandelbaum et al. (in preparation). In the first condition we asked participants if an agent is responsible for actions that are determined by a neurological illness outside of his control. Here, the participants overwhelmingly did not deem the agent responsible (mean response: 2.6). Importantly, in this first condition we did not specifically mention what action the agent performed; rather we left the action only abstractly described as "a certain behavior." In the second condition, the agent is in a similar neurological situation, but we mention a particular bad action (rape). Once the action was described in concrete terms, participants held the agent significantly more responsible (mean response: 4.2). This is further evidence, in line with Nichols and Knobe's study, that folk judgments of responsibility are strongly affected by the specificity and affective nature of the vignettes presented.

Here is a tentative suggestion about these results: when faced with a concrete bad action, participants experience an impulse to blame someone for that bad action (see e.g. Alicke 2000; Nadelhoffer 2006 for similar hypotheses). This sort of hypothesis is compatible with many different theories of the etiology of the impulse itself. One might hold that the impulse is an emotional affair, like the "affective bias" proposed by Nichols and Knobe to explain similar effects, or one might hold that the impulse is cognitively mediated, like the cognitive dissonance effects proposed by e.g. Maikovich (2005) to explain other cases of seemingly unreflective blame. Of course, one might think, as we do, that both cognitive and affective factors are at work as well; hopefully, future research will shed light on the causes and structure of the impulse to blame.

If this is right, what does it mean for the courtrooms of the future? We think Greene and Cohen's picture underestimates the folk's tendency to hold people responsible as a result of specificity and affect. As neuroscience evolves in the laboratory, its explanations will not simply be taken over into society at large; rather, its results will be interpreted and incorporated into our larger worldview. In particular, as Greene and Cohen realize, neuroscience presumably will yield no results directly about responsibility, but instead will at most undercut the folk's intuitive dualism. Now, as Haidt (2001) has stressed, our moral worldview is socially negotiated, partly in response to our quick and intuitive judgments. But these intuitive judgments are subject to the well known anchoring and adjustment phenomenon (e.g. Tversky and Kahneman 1974), which should make us a bit skeptical about Greene and Cohen's wide-sweeping claims about the juries of the future. People may move away from their intuitive judgments, but their intuitive judgments serve as an anchor; moving away from these judgments is always an uphill battle. Thus, prima facie, we should expect that it is unlikely that our future neuroscience will be able to completely overtake such well-entrenched intuitive judgments. Pace Greene and Cohen, we have some reason to expect a future in which juries both a) see that defendants' behaviors are strictly determined by their neurological states and yet b) still find that defendants have enough free will to be held responsible.[10]

[10] Presumably, whether this type of future is seen in a positive or negative light will depend on, *inter alia*, whether one is a compatibilist or an incompatibilist. The authors of this paper are currently split on whether or not such a future is desirable. We thank an anonymous referee for helping us bring out this point.

6 Objections and Replies[11]

Objection The survey method doesn't tell us about what judgments people will actually make in everyday life.

Reply Survey methods, like any other methodology, are only one of multiple tools that can be useful in shedding light on these very difficult questions. However, we do think that this method is reputable: first, the method is often used by social scientists (following e.g. Likert 1932), and has been adapted with success by experimental philosophers (e.g. Knobe 2003). Second, in this particular case, the surveys seem to accurately track the extant data pertaining to insanity defenses. In the United States, insanity defenses except from criminal responsibility people who have a disability which impairs either their cognitive or volitional capacities. Thus, our characters are very much like defendants who are able to mount an insanity defense. If our survey data is accurate, we'd predict that the majority of people will still hold defendants with cognitive or volitional deficits culpable for their behavior, and thus will find defendants guilty as opposed to innocent. As we'd predict, the insanity defense works astonishingly poorly. Insanity defenses are successful only a little more than 15% of the time (Borum and Fulero 1999).[12] Thus, the actual data on insanity defenses lines up quite nicely with our studies.

Objection Surveys don't allow us to draw any conclusions about the mechanisms by which participants arrive at their judgments.

Reply Theoretical models are always underdetermined by empirical evidence (see e.g. Gilbert 1999). However, we do not think that we are in any worse a position for having a survey-based methodology than any other methodology—this is the position that social scientists often find themselves in. We, of course, do not think that the question of what mechanisms are causing our participants' behavior is anywhere near settled. But our studies do rule out some models—particularly, they rule out a model that has been quite influential: one where facts about psychology are processed differently than facts about neurology, come what may. We may not be able to show exactly how the judgments are made, but we can be sure about how they are not made, which is progress.

Yet we think that we are entitled to an even stronger conclusion. Since models of what's going on inside a person's head are always underdetermined by data, whether that data is survey based or neuro-image based, one way of deciding between competing models is by seeing what machinery the model requires. Our explanation relies only on very well-grounded and well-studied cognitive processes: dissonance theory (e.g. Festinger 1957) and affective biases (e.g. Lerner et al. 1998). These are two of the bedrock forms of explanations alive in the cognitive sciences. Although we can't be sure that our explanation is correct, it should at least be given some preference because it relies on such well-entrenched principles.[13]

[11] We thank an anonymous referee for raising the concerns considered in this section.

[12] Moreover, their lack of success rate is probably not due to defendants transparently pretending to be insane. If this were the case, we'd expect a high rate of felony defenses to be predicated on the insanity defense; in fact, the defense is raised less than 1% of the time, presumably because lawyers know how poorly the defense works (ibid.).

[13] How the actual mechanisms work is the focus of our next paper, "Abstract Thought in Concrete Situations" (Mandelburn et al. in preparation).

Lastly, our model for explaining how the folk make their judgments is empirically testable in many ways. Perhaps the most salient way would be to ask participants to read these vignettes and respond while having their brains scanned in an fMRI machine. If we are right that there is an affective bias at play, then the neurological regions which underwrite affect should be active while the participants are reading and responding to our surveys. Specifically, we predict that we will see more activation in the neurological regions associated with affect in our participants when they read the concrete vignettes in our first three studies than when they read the abstract vignettes in Nahmias's studies. We also predict higher activation in these regions when participants read the concrete vignette in our follow-up study (see Section 5) than when they read the abstract vignette.

In sum, even though we agree that surveys cannot tell us exactly how the mind works, they do provide some evidence. Furthermore, our explanation of why this effect arises is both empirically falsifiable and relies on only well-known psychological principles.

7 Final Remarks

Greene and Cohen's most persuasive argument for their claims about the folk of the future involves a character named Mr. Puppet. Mr. Puppet would not exist at all except for the dastardly plans of devious scientists, who carefully selected each of Mr. Puppet's genes, and raised him in a carefully controlled environment, making sure to traumatize him in very precise ways. All of this eventually causes Mr. Puppet to commit a murder, just as the scientists had planned. Greene and Cohen claim that, intuitively, Mr. Puppet is not responsible for the murder he commits. They also claim that to the folk of the future, we will all look roughly like Mr. Puppet does to us now; after all, all our behaviors are, like his, the product of some (as-yet-ill-understood) combination of our genes and our environment. Thus, to the folk of the future, nobody will seem intuitively responsible.

We think this is too quick. There is a difference between Mr. Puppet and the rest of us: in Mr. Puppet's case, *there is someone else to blame*. As far as present intuition is concerned, Mr. Puppet isn't responsible for the murder, but the scientists are. For juries of the future, though, there will be no dastardly scientists to blame.[14] Because of this, we think that current intuitions about Fred, Cliff, Jennifer, Dennis, and Jerry are more likely to match these future intuitions about normal action than are current intuitions about Mr. Puppet. Our five agents have their actions caused either by their illnesses or by distractions; there is no other agent on whom to displace the blame. Since at present, the folk seem to judge that these five are somewhat responsible, we have some evidence that the folk of the future will still hold each other somewhat responsible, even if they find that each other's actions are caused in surprising ways.

Acknowledgments Many thanks to the audience at the *Society for Philosophy and Psychology* at York University in Toronto, Canada, the audience at the 2008 Central Division Meeting of the APA in Chicago, IL, the audience at *Mind, Brain, and Experience* at the University of Colorado in Denver, and the audience at the tenth anniversary conference for the journal *Ethical Theory and Moral Practice* at the Blaise Pascal Institute in Amsterdam, Netherlands. Many thanks also to Bryce Huebner, Mark Phelan, and Jesse Prinz for insightful comments, and to Joshua Knobe for his unprecedented generosity. Thanks also to two anonymous referees for their helpful comments.

[14] There may well be e.g. dastardly parents or dastardly societal forces to blame; to the extent that these are present and well-understood, we might expect to see lowered responsibility judgments. But neurology alone will not put every defendant in this situation; even a complete understanding of a behavior's neural causes could leave open the story of how the neurons came to be the way they are.

References

Alicke M (2000) Culpable control and the psychology of blame. Psychol Bull 126(4):556–574, doi:10.1037/0033-2909.126.4.556

Borum R, Fulero S (1999) Empirical research on the insanity defense and attempted reforms: evidence towards informed policy. Law Hum Behav 23(1):117–135, doi:10.1023/A:1022330908350

Festinger L (1957) A theory of cognitive dissonance. Stanford University Press, Stanford

Fodor J (1974) Special sciences (or the disunity of science as a working hypothesis). Synthese 28(2):97–115, doi:10.1007/BF00485230

Gazzaniga M (2005) Neuroscience and the law. Sci Am Mind 16(1):42–49

Gilbert D (1999) What the mind's not. In: Chaiken S, Trope Y (eds) Dual process theories in social psychology. Guilford, New York, pp 3–11

Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. Philos Trans R Soc Lond B Biol Sci 359:1775–1785 doi:10.1098/rstb.2004.1546

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgement. Psychol Rev 108:814–834

Knobe J (2003) Intentional action and side effects in ordinary language. Analysis 63:190–193, doi:10.1111/1467-8284.00419

Lerner J, Goldberg J, Tetlock P (1998) Sober second thought: the effects of accountability, anger, and authoritarianism on attributions of responsibility. Pers Soc Psychol Bull 24(6):563–574 doi:10.1177/0146167298246001

Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 140:1–55

Maikovich A (2005) A new understanding of terrorism using cognitive dissonance principles. J Theory Soc Behav 35(4):373–397, doi:10.1111/j.1468-5914.2005.00282.x

Mandelbaum E, Ripley D, De Brigard F (in preparation) Abstract thought in concrete situations

Nadelhoffer T (2006) Bad acts, blameworthy agents, and intentional actions: some problems for jury impartiality. Philos Explor 9(2):203–220, doi:10.1080/13869790600641905

Nahmias E (2006) Folk fears about freedom and responsibility: determinism vs. reductionism. J Cogn Cult 6:215–237, doi:10.1163/156853706776931295

Nichols S, Knobe J (2007) Moral responsibility and determinism: the cognitive science of folk intuitions. Nous 41(4):663–685, doi:10.1111/j.1468-0068.2007.00666.x

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Sci 185:1124–1131, doi:10.1126/science.185.4157.1124