Forthcoming in *Journal of Philosophy*

Bringing moral responsibility down to earth[*]

Thought experiments have played a central role in philosophical methodology, largely as a means of elucidating the nature of our concepts and the implications of our theories.[1] Particular attention is given to widely shared "folk" intuitions – the basic untutored intuitions that the layperson has about philosophical questions.[2] The folk intuition is meant to underlie our core metaphysical concepts, and philosophical analysis is meant to explicate or sometimes refine these naïve concepts. Consistency with the deliverances of folk intuitions is a sign that the philosopher is making contact with his object of interest. In order to explore folk concepts, people are often asked to provide their intuitions about a state of affairs in some alternate universe or possible world, one that differs in particular, precise ways from the way things are in the actual world. Here we provide evidence that people's intuitions about moral responsibility sometimes diverge across worlds even when the facts about these worlds are the same. Which world one considers actual affects at least some philosophical judgments, suggesting that it is not just possible worlds to which our intuitions are tied. We will present several possible explanations for the asymmetry we have identified, and we'll consider some implications for philosophical intuition.

It has frequently been claimed that the folk are incompatibilists about freedom and moral responsibility – that they believe that freedom is not possible in a deterministic universe, and that if you are not free, you are not morally responsible.[3] Thus Kane claims, "In my experience, most ordinary persons start out as natural incompatibilists."[4] And Pereboom writes, "Beginning students typically recoil at the compatibilist response to the problem of moral responsibility." [5] Of course, other philosophers have suggested that the common view is actually compatibilist.[6]

The nature of our intuitions about free will and moral responsibility is not, however, purely a matter of a priori debate. What the folk think is an empirical question, and one which can be addressed by

[*] The authors contributed equally to this work. We received helpful comments on this work from a number of people. We would like especially to thank David Braddon-Mitchell, John Fischer, Richard Holton, Joshua Knobe, Uriah Kriegel, Derk Pereboom, Walter Sinnott-Armstrong, Saul Smilansky, and Roy Sorensen.

[1] Roy Sorensen, *Thought Experiments* (Oxford: Oxford University Press, 1992).

[2] Frank Jackson, *From metaphysics to ethics: A defence of conceptual analysis.* (Oxford: Oxford University Press, 1998).

[3] As John Fischer has emphasized in "Recent Work on Moral Responsibility" *Ethics*, CX (1999)it is possible that determinism is consistent with responsibility but not with free will. Thus, it is often important to distinguish between *moral-responsibility compatibilism* and *free will compatibilism,* and we will keep the views distinct here.

[4] R. Kane "Responsibility, luck, and chance: Reflections on free will and indeterminism" THIS JOURNAL, XCVI (1999): 217.

[5] Derk Pereboom, *Living Without Free Will* (Cambridge: Cambridge University Press, 2001) p.xvi.

[6] For example, David Hume's *Enquiry concerning Human Understanding* (1748).

what is coming to be known as "experimental philosophy". Recently, experimental philosophical approaches have yielded conflicting results as to whether the folk are, in general, compatibilists or incompatibilists about moral responsibility. Nahmias and colleagues had subjects assume that determinism is true, and then judge whether an agent is blameworthy under those circumstances. They found that subjects tended to say that the agent was blameworthy.[7] Using a different experimental design, Nichols & Knobe (2007) presented subjects with a description of an alternate universe that is deterministic, and they found that subjects tended to say that agents were not responsible in that universe.[8]

This pattern of conflicting results suggests that subtle features about the way questions about moral responsibility are framed may have an effect upon our intuitions. There are a number of differences between the experiments done by Nahmias et al. and those done by Nichols and Knobe. But for present purposes, we are interested in just one of these differences. In some of the experiments by Nahmias et al., the scenario was depicted as holding of our own world, whereas in the experiments by Nichols & Knobe, the scenarios were always set in an alternate universe. For our experiment here, we wanted to see what would happen if we posed questions that differ only in whether the hypothetical situation was set in our own universe or another. On the face of it, one would expect that responses to a hypothetical situation would depend solely upon features of that situation, and not upon the subject's relation to that situation. Thus, one might expect that responses to questions about freedom and responsibility would be independent of whether those questions were couched in terms of our universe, or another universe just like ours. In the following experiment, we tested this expectation.

Down to earth

Our study was conducted on University of Utah undergraduates who hadn't been exposed to philosophical instruction about free will or responsibility.[9] This experiment depends upon our ability to convey, in layman's terms, the essential features of determinism. Although the most precise characterizations of determinism depend on technical philosophical vocabulary about, for instance, laws of nature, we were interested in asking philosophically naïve subjects about their intuitions. Thus, it was important to translate some technical terms into familiar nonphilosophical language. Different characterizations of determinism might produce somewhat different results, but our central interest concerns whether different conditions generate different responses *given the same description of determinism.* One of the great advantages of contrastive studies is that they can be designed to determine whether one factor makes a difference to a response. The factor of interest to us is whether the scenario is set in the actual world or an alternate universe.

---

[7] E. Nahmias, S. Morris, T. Nadelhoffer, & J. Turner "Surveying Freedom: Folk intuitions about free will and responsibility" *Philosophical Psychology*, XVIII,5 (2005): 561-584.

[8] S. Nichols and J. Knobe "Moral responsibility and determinism: The cognitive science of folk intuitions" *Nous*, XLI, 4 (2007): 663-685.

[9] We hope to explore whether these results extend to other populations. (We hope even more that *others* will explore whether our results extend to other populations!) However, it's important to note that we are primarily looking at whether subjects from the same population give different answers in the different conditions.

76 participants were randomly assigned to one of two conditions: *Actual* and *Alternate*. In both conditions, subjects were given the same sketch of a universe that is regarded as deterministic. In the *Actual* condition, this universe was clearly implied to be our own:

> 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 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.

In the *Alternate* condition, the universe is explicitly not ours:

> 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, 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 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.

In both conditions, participants were then asked to rate their level of agreement (from 1 [disagree completely] to 7 [agree completely]) with a statement about the impossibility of moral responsibility in the universe. In the *Alternate* condition, they were presented with this statement:

> If these scientists are right, then it is impossible for a person in Universe A to be fully morally responsible for their actions.

In the *Actual* condition, participants received the same statement, except that "in Universe A" was removed.

> If these scientists are right, then it is impossible for a person to be fully morally responsible for their actions.

The purpose of this experiment was to determine whether judgments would be affected by whether the question was asked of this world or some other possible world. The results were striking: participants in the *Alternate* condition gave significantly higher levels of agreement than participants in the *Actual* condition to the claim that it is impossible to be fully morally responsible in that

universe.[10] So, when asked to assume that our own universe is deterministic, people are inclined to judge that people are still morally responsible (in other words, they provide compatibilist responses), but they are inclined to judge of people in another deterministic universe that they are not fully morally responsible, an incompatibilist response.[11]

After answering the question about the impossibility of moral responsibility, participants were asked to indicate their level of agreement with a statement about the moral propriety of blame in the deterministic universe. In the *alternate* condition, they received the following statement:

> Even if these scientists are right, people in Universe A should still be morally blamed for committing crimes.

In the *actual* condition, participants received the same statement, but "in Universe A" was removed:

> Even if these scientists are right, people should still be morally blamed for committing crimes.

Once again, responses were quite different in the different conditions. Participants in the *actual* condition gave significantly higher levels of agreement to this claim. That is, they were significantly more likely than those in the *alternate* condition to give compatibilist responses to this question as well.[12]

Finally, participants were also asked to indicate their level of agreement with a statement about the existence of free choice in the deterministic universe. In the *Alternate* universe, they received the following statement:

> If these scientists are right, then it is impossible for people in Universe A to make truly free choices.

In the *actual* condition, participants received this:

> If these scientists are right, then it is impossible for people to make truly free choices.

---

[10] The mean response in *Actual* was 3.58, and the mean response in *Alternate* was 5.06 (4 is the midline). The difference between the conditions was significant (t (74) = 3.611, p =.001).

[11] Interestingly, these responses parallel Peter van Inwagen's claim that he thinks that moral responsibility is incompatible with determinism but that if he came to believe determinism was true, he would be a compatibilist. (see *An essay on free will* (Oxford: Clarendon Press, 1983)). Fisher & Ravizza label this view "metaphysical flipflopping" in *Responsibility and Control : A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998).

[12] The mean response in *Actual* was 5.35, and the mean response in *Alternate* was 3.67. The difference between the conditions was significant (t (65) = -4.426, p <.001). Note that the question was positively framed, while the other two were framed negatively, so we in fact would expect this sort of inversion in the agreement values. These values are consistent with what would be expected if judgments of moral responsibility, blameworthiness, and freedom were correlated.

For this question as well, there was a significant difference between conditions. Participants in the *Alternate* condition gave higher ratings of agreement to this statement than those in the *Actual* condition.[13]

<insert figure here>

Implications for our understanding of moral responsibility

The above results have several implications for our understanding of moral responsibility. First, they provide one way of interpreting some of the apparently conflicting findings of Nichols and Knobe and of Nahmias and colleagues.[14] For we find that subtle differences in the way questions about moral responsibility are framed can make a difference to the intuitions that are elicited. In some of the scenarios from Nahmias and colleagues, subjects are explicitly asked to assume that determinism is true of *our universe*, whereas in Nichols & Knobe's scenarios, subjects are always asked to assume that determinism is true of an alternate universe. Thus, our results suggest that part of the explanation for the diverging responses might be the setting of the scenario. When the deterministic universe is our own, people are more likely to give compatibilist responses.

Secondly, the results bear upon an issue that has captured both the philosophical and the popular imagination. Some philosophers and many laypeople fear that catastrophe will follow if people come to accept determinism. For example, Saul Smilansky suggests that if people come to realize the absence of indeterminist choice, "our fundamental values, practices, and attitudes, such as abhorrence about the 'punishment' of the innocent, the inherent value we put on 'equality of opportunity', belief in our potential for blameworthiness… are very likely to be harmed."[15] Smilansky worries that people "might succumb to … an unprincipled nihilism."[16] As a result, although Smilansky thinks that free will is an illusion, he maintains that we should not try to dispel the illusion. Related claims have been made in the popular press in the context of neuroethics. Among the neuroethical worries raised by technological advances in neuroscience is that our improving scientific understanding of higher brain functions will cause the public to view currently unexplained psychological phenomena such as choice and decision-making as a merely mechanical processes, and that they will come to believe that human action is merely the result of mechanism in a deterministic universe.[17] The projected upshot of this potential change in belief is that upon coming to see us as merely mechanisms in a deterministic world we will come to realize that we lack free will, and consequently, moral responsibility. This in turn, the argument continues, will undermine the moral fabric of

---

[13] While the difference is significant (t (74)=2.362, p <.05), in both conditions, the mean response was above the midline. In the *Actual* condition, the mean response was 4.3, and in the *Alternate* condition, the mean response was 5.3. Overall, these results suggest that there was stronger agreement to the statement that people (in a determinist universe) lacked freedom of choice than that they lacked moral responsibility. Interestingly, we found that this difference was statistically significant in the *Actual* condition (t (39) = -2.456, p < .05), but not in the *Alternate* condition (t (35) = -.843, p = .405, n.s.).

[14] S. Nichols and J. Knobe (Op. cit.), and E. Nahmias, S. Morris, T. Nadelhoffer and J. Turner (Op. cit.).

[15] S. Smilansky *Free Will and Illusion* (Oxford: Oxford University Press, 2000) p. 189.

[16] Ibid. p. 189.

[17] For a discussion of why such a conclusion is not warranted on the basis of neuroscientific research see A.L. Roskies "Neuroscientific challenges to free will and responsibility" *Trends in Cognitive Sciences*, X (2006): 419-423.

our society; the chaos that will result is left to our imaginations. [18]

This line of reasoning is often meant to cast doubt on the ethics of pursuing neuroscientific research into the brain bases of higher cognition. This ploy is questionable for a number of reasons. What is crucial for our purposes, however, is whether the folk will stop treating people as morally responsible if determinism becomes widely accepted. Our experiment addresses this concern in two ways. First, it provides some explanation for the expectation that the belief in determinism would lead us to abandon our belief in moral responsibility. For when asked about a hypothetical situation in which determinism is true (the Alternate case), people are inclined to claim that moral responsibility is excluded. More importantly, though, the experiment suggests that the practical worries are misplaced, for our judgments about moral responsibility, should we come to believe determinism to be true of the *actual* world, would probably not be unseated by this belief.[19] The upshot of this is that these worries about how neuroscientific understanding will undermine the social order are misplaced.[20]

Other philosophers agree with Smilansky that people's attitudes towards responsibility will change markedly if they come to believe in determinism, but instead they hail this as a much needed revolution. Thus, Joshua Greene and Jonathan Cohen write, "As more and more scientific facts come in, providing increasingly vivid illustrations of what the human mind is really like, more and more people will develop moral intuitions that are at odds with our current moral practices." [21] In particular, they maintain, we will stop thinking that people are responsible and that the guilty deserve punishment. "The law will continue to punish misdeeds, as it must for practical reasons, but the idea of distinguishing the truly, deeply guilty from those who are merely victims of neuronal circumstances will… seem pointless."[22] Our results suggest that we should expect neither revolution, as do Greene and Cohen, nor catastrophe, as do Smilansky and the naysayers in the popular press. If people came to believe in determinism, it seems likely that they would not significantly change their practices of attributing responsibility.

---

[18] See, for example, F. Fukuyama *Our posthuman future: Consequences of the biotechnology revolution*. (New York: Farrar, Straus & Giroux, 2002); C. Goldberg. "A question of the will\" *The Boston Globe* (October 15, 2002) p. C1; R.J. Rychlak and J.F. Rychlak "Free will is a verifiable assumption: a reply to Garrett and Viney" *New Ideas in Psychology*, VIII (1990) p. 43-51; and T. Wolfe, (1996). "Sorry, but your soul just died" *Forbes*, CLVIII (1990) p. 210. There are a variety of deterministic accounts of decision and action. Most work in neuroethics does not clearly distinguish between neurological determinism and metaphysical determinism. For some purposes it matters whether the deterministic thesis is neurological, but for our purposes it does not. For ease of exposition we will talk about determinism generically.

[19] For arguments to a similar conclusion, see T. Nadelhoffer and A. Feltz "Folk Intuitions, Slippery Slopes, and Necessary Fictions: An Essay on Saul Smilansky's Illusionism" *Midwest Studies in Philosophy*, XXXI (2007) p. 202-213.

[20] This is not to say that neuroscientific advances pose no threat to the social order; only that this particular focus for worry is overblown.

[21] J. Greene and J.D. Cohen " For the law, neuroscience changes nothing and everything" *Phil. Trans, R. Soc. Lond. B*, CCCLIX (2004) p. 1781.

[22] Ibid.

What is going on?

Our results are surprising – why should intuitions about moral responsibility depend upon factors other than the basic facts about the world? There are several models by which to understand these findings.

Depth of processing

One explanation of our results appeals to differences in the depth of processing in the different conditions. It is an old and venerable view in social psychology that people process problems more fully and accurately when the questions are personally relevant.[23]. For instance, in one study, undergraduate participants were presented with arguments advocating decreased privileges in dorms. One group (high involvement) was told that the plan was for their own university; the other group (low involvement) was told that the plan was for another university. Those in the highly involved group showed a greater sensitivity to the quality of the arguments than those in the other group.[24] Obviously in our experiments, the *Actual* condition has greater personal relevance for participants than the *Alternate* condition. So perhaps this is why participants give different answers in the different conditions.

Motivation

A different explanation for our results appeals to the role of motivational factors. Social psychologists have long known that people are more likely to believe things they want to be true than things they don't want to be true.[25] There are a number of different ways in which motivation influences belief. For instance, evidence that supports an unwanted conclusion is less likely to be remembered or trusted than evidence that supports a desirable conclusion.[26] In the present case, we *want* to hold people responsible in our world, but we are less motivated to hold people responsible in an alternate universe. So this would provide participants with additional inclination to give the compatibilist response in the *Actual* condition.

---

[23] See, for example, J.A. Bargh "Automatic and conscious processing of social information" In R. S. J. Wyer & T. K. Srull (Eds.), *Handbook of social cognition,* III, (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc, 1984) p.1-44. See also S. Chaiken, "Heuristic versus systematic information processing and the use of source versus message cues in persuasion" *Journal of Personality and Social Psychology*, XXXIX (1980) p. 752-766;, and S. Chaiken, A. Liberman and A. Eagly "Heuristic and Systematic Information Processing Within & Beyond the Persuasion Context" in J. S. Uleman & J. A. Bargh (Eds.), *Unintended Thought* (New York: Guilford Press ,1989) pp. 212-252, and R. Petty and J. Cacioppo "The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion." *Journal of Personality and Social Psychology*, XLVI (1984) p. 69-81.

[24] R. Petty and J. Cacioppo "Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses" *Journal of Personality and Social Psychology*, XXXVII, (1979) p. 1915-1926.

[25] Z. Kunda, *Social Cognition* (Cambridge: MIT Press: 1999).

[26] B. Doosje, R. Spears, and W. Koomen "When bad isn't all bad: Strategic use of sample information in generalization and stereotyping" *Journal of Personality and Social Psychology*, LXIX (1995) p. 642-655; R. Sanitoso, Z. Kunda and G. Fong, "Motivated recruitment of autobiographical memories" *Journal of Personality and Social Psychology*, LVII (1990) p. 229-241.

A possibly more troubling motivational explanation appeals to dissonance reduction.[27] We routinely hold people responsible for their actions, so when subjects take seriously the prospect of determinism, this will create cognitive dissonance between their past practices and their incompatibilist intuitions. According to dissonance theory, the tension is relieved by bringing their views about responsibility in line with how they have practiced the attribution of responsibility. That is, people reduce the dissonance by moving to more compatibilist views of responsibility.[28]

A related explanation makes reference to what Frank Jackson terms a "non-robust conditional".[29] A non-robust conditional is one that a person accepts, but would reject if sufficient evidence were to convince him that the antecedent was true. In this case, it seems that people believe the non-robust conditional "If the universe is deterministic, then we are not morally responsible for our actions." But, given the antecedent, people are inclined to reject the conditional rather than accept its consequent, that we are not morally responsible for our actions. The explanation in terms of non-robust conditionals may be a version of a cognitive dissonance story, for some non-robust conditionals are presumably nonrobust because embracing the consequent is so costly. However, the appeal to non-robust conditionals differs from a standard cognitive dissonance story in that it does not imply that people are being irrational in rejecting the conditional they previously endorsed. Rather, they are justified in rejecting it, for the failure lies in the conditional itself, and not the believer. As Sorensen puts it, "some knowledge of conditionals is "junk knowledge"; one cannot go from knowledge of the conditional to knowledge of the consequent."[30]

Affect

A further psychological interpretation of our results can be gleaned from moral psychology and neuroscience. Although moral deliberation involves reasoning and sophisticated cognitive processes, a number of studies have shown that brain areas involved in emotional processing are also recruited in the making of some moral judgments.[31] In addition, in cases in which emotional regions are recruited, subjects are more inclined to judge a scenario as morally wrong or a person as morally

---

[27] L. Festinger *A theory of cognitive dissonance* (Stanford, CA: Stanford University Press, 1957).

[28] We thank Richard Holton for suggesting the cognitive dissonance explanation to us.

[29] F. Jackson *Conditionals* (Oxford: Oxford University Press, 1991).

[30] R. Sorensen "Dogmatism, Junk Knowledge, and Conditionals" *Philosophical Quarterly*, XXXVIII (1988) p. 433- 454.A non-motivational interpretation of the results may also be closely related to the non-robust conditional story. Just this pattern holds in cases in which one's credence in the conditional is less strong than one's credence in the consequent in some contexts. Forced to give up something in those contexts, one will give up the conditional. On this view, the explanation wouldn't be motiviational, but rather cognitive. We thank Andy Egan for pointing this out.

[31] J.D. Greene, R.B. Sommerville, L.E. Nystrom, J.M. Darley and J.D. Cohen "An fMRI investigation of emotional engagement in moral judgment" *Science*, CCXCIII, (2001) p. 2105–2108; J. Moll, R. de Oliveriera-Souza, P.J. Eslinger, I.E. Bramati, J. Mourao-Miranda, P.A. Andreiuolo, et al. "The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions" *Journal of Neuroscience*, XXII (2002) p. 2730-2736; and J. Schaich Borg, C. Hynes, J. Van Horn, J. S.T. Grafton and W. Sinnott-Armstrong, "Consequences, action and intention as factors in moral judgments: An fMRI investigation" *Journal of Cognitive Neuroscience*, XVIII, (2006) p. 803-817.

blameworthy than in cases in which affect is not a major factor. In fact, people's judgments of moral turpitude or blameworthiness can be manipulated by manipulating their negative emotions of anger or disgust: Haidt and colleagues have shown that people who score are especially attuned to their bodily sensations (as measured by the "Private Body Consciousness" scale) report greater moral condemnation concerning a vignette when they are sitting at a disgustingly dirty desk than when they are sitting at a clean desk.[32] In addition, people's moral judgments can be modulated by hypnotically induced disgust.[33] Thus, a growing number of experiments from both psychology and neuroscience indicate that brain areas involved in emotional processing are involved in some moral judgments, and that emotion is causally efficacious in modulating judgments of responsibility and wrongness.

Further substantiating results come from Nichols and Knobe's experiments on determinism and responsibility. In their experiments, people responded in ways that suggested that they are natural incompatibilists: they think that we live in an indeterministic universe, and that people in a determinist universe would not be fully morally responsible. Indeed, it is only on this view that one would think it plausible that coming to believe that the universe is deterministic would undermine our notions of free will and responsibility. However, their experiments also demonstrated that when given concrete scenarios in which an actor operated in a deterministic universe, people's judgments of moral responsibility tracked their level of affect. In concrete cases that were unlikely to trigger much affect, people tended to say that the actor could not be morally responsible for his actions; as the scenario was described in more detail or using more affect-arousing language, people judged that the actor could be morally responsible.

Our results could be understood as a consequence of the variable involvement of emotion in the assessment of scenarios set in our own or in other worlds. One can think of alternate universes as more removed and less personally involving than our own, so that the very same scenarios would differentially involve emotional areas during processing of questions of moral responsibility. This differential involvement would explain the differences we see in judgments of moral responsibility across worlds.

Each of the three preceding accounts – depth of processing, motivation, and affect – is a psychological explanation of the results. However, we should note that these three explanations are not mutually exclusive. It's quite possible that the right psychological explanation is a hybrid. For instance, one live option is that subjects respond as compatibilists in the *Actual* condition as a result of deeper processing, which itself is a product of motivation. Indeed, one way to improve subjects' performance in evaluating evidence is by making the putative conclusion of the evidence loathe to the subjects. In one study, subjects were given a data set that purported to show that women had inferior leadership skills, but careful inspection reveals that the data do not support the conclusion; the female subjects, who were highly motivated to reject the conclusion, were more likely to discern that the data did not support the conclusion.[34] Similarly, then, perhaps subjects in the actual

---

[32] S. Schnall, J. Haidt, G. Cloreand A. Jordan (forthcoming). "Disgust as embodied moral judgment" *Personality and Social Psychology Bulletin*.

[33] T. Wheatley and J. Haidt "Hypnotically induced disgust makes moral judgments more severe" *Psychological Science*, XVI (2005) p. 780-784.

[34] M. Schaller "In-group favoritism and statistical reasoning in social inference" *Journal of Personality and Social*

condition in our experiment were motivated to think through the problem more carefully than subjects in the alternate condition. It is likely that further experimental investigation into this phenomenon will enable us to rule out one or more of the potential psychological explanations.

Under most of the foregoing psychological explanations, people are making some kind of mistake or cognitive error. On the cognitive dissonance proposal, there is a presumption of irrationality in the reduction of the dissonance; on the non-robust conditional account, subjects realize they were mistaken in accepting the conditional initially, upon learning that its antecedent is satisfied; on affect-based accounts, subjects can be construed as falling prey to a performance error when their emotions are activated. There are, however, other, non-psychological, interpretations of subjects' responses that do not imply that subjects are mistaken in their judgments. We now turn to one such interpretation.

Analytical conditional analyses

Our results could be understood on a model that draws from analytic metaphysics rather than psychology. The results we find here should be quite congenial, and maybe even encouraging to theorists inclined toward two-dimensional semantics. We draw upon an analysis by David Braddon-Mitchell to illustrate.

Braddon-Mitchell has recently provided a conditional analysis of qualia, based on a 2-dimensional (2D) approach, in order to explain a perennial puzzle in the philosophy of mind.[35] The dilemma he responded to is this: How can we believe at the same time that (a) the world is purely physicalistic, so there is no spooky stuff; and (b) that it is nonetheless conceivable that there are physical duplicates of us that lack qualitative states or qualia? In other words, how it is that we can be at the same time physicalists and countenance the possibility of zombies? Braddon-Mitchell's conditional analysis provides an analysis of the concept of qualia (henceforth <quale>) that provides us with a coherent account of both intuitions. He argues that the correct account of <quale> is:

> If there are spooky states then in the actual world the qualia are the spooky states, and all and only the qualia in counterfactual worlds are the spooky states; else, in the actual world the qualia are the states that play the functional roles [we take qualia to play]; in other worlds, qualia (if any) are the states that play those roles *in that world*. (Braddon-Mitchell, 2003, p. 121)[36]

In order to explicate his account, Braddon-Mitchell uses the framework provided by 2-dimensional semantics. This framework evaluates sentences along two dimensions, each of which can be represented by sets of possible worlds. The sentence's truth value is dependent upon both the

---

Psychology, LXIII (1992) p. 61-74.

[35] D. Braddon-Mitchell "Qualia and Analytic Conditionals" This Journal, C (2003) p. 111–135.

[36] In "Qualia and Analytical Conditionals" Braddon-Mitchell (*Op. cit.)* goes on to refine this view taking account of the importance of centered possible worlds; this is a refinement that doesn't matter much for our purposes, since our judgments about moral responsibility are not made on the basis of first-person experiences.

features of the possible world at which it is evaluated, and also upon which world is taken as actual. Here we will follow the analysis of Jackson.[37]

The C-intension is familiar – it's just the set of truth values of the sentence evaluated at every possible world, taking one particular world as fixed. (So, for instance, taking our world as actual, the C-intension of the sentence "Swimming pools often have water in them" would be assigned the truth value *false* at Twin Earth, because "water" refers to H2O, and there isn't any H2O in that world.) By constrast, the A-intension of a sentence is the function which evaluates the truth value of the sentence at every world taking that world as actual. (So, for instance, the A-intension of "Swimming pools often have water in them" would be assigned the value *true* at Twin Earth, because XYZ is "water" on Twin Earth.)[38] Braddon-Mitchell uses 2D semantics to provide a principled grounding for the conditional analysis, and to derive a number of conclusions about possibility and necessity claims relating to qualia.[39] Of particular interest is that the A-intension of "there is a physical duplicate of me that lacks qualia" is contingent (or true at some worlds and false at others), for in worlds taken as actual in which spooky states exist, physical duplicates lack qualia, and in physicalist worlds taken as actual, they have them. This contingency accounts for the sense in which zombies are possible. However, in physicalist worlds taken as actual, the C-intensions of "there is a physical duplicate of me that lacks qualia" are necessarily false (false at every world), thereby accounting for the physicalist intuition that zombies cannot exist. Thus, the 2D analysis explains how we can reconcile the conceivability of physical duplicates without qualia (zombies) with our physicalist intuitions. Moreover, Braddon-Mitchell argues, the conditional analysis of <quale> itself has a necessary A-intension, grounding the intuition that it provides an analysis of our concept.[40]

---

[37] F. Jackson, 1998, *Op. cit.*

[38] This may be more easily grasped by thinking of the 2D analysis in terms of matrices, the cells of which represent the truth value of the sentence evaluated under two conditions, corresponding to the possible world at which it is evaluated (columns), and the world which is taken as actual (rows). Each cell in the matrix is indexed by row and column entries denoted by the ordered pair *(i,j)*. The row entries *(i)* of the matrix represent various ways our world could be, and the columns *(j)* represent various features of possible worlds at which the sentence is to be evaluated. The entry at matrix cell *(i,j)* provides the truth value of the sentence in question evaluated at world *j* taking world *i* as actual. The A-intension and the C-intension are then understood geometrically: The A-intension is represented by the diagonal of the matrix (entries where $i = j$), and the rows of the matrix give the various C-intensions of the sentence. There are as many C-intensions as there are ways of conceiving of the world as actual. For a detailed explication of the 2-D analysis we rely upon, see Jackson (1998, Op. cit.) and Braddon-Mitchell (Op. cit.). David Chalmers has a similar analysis in *The Conscious Mind: In search of a fundamental theory* (Oxford: Oxford University Press, 1996). Other two-dimensionalist accounts vary on a number of dimensions; for a state-of-the-art anthology see M. Garcia-Carpintero and J. Macia, (Eds.). *Two-Dimensional Semantics* (Oxford: Oxford University Press, 2006).

[39] Braddon-Mitchell, *Op. cit.*

[40] *Ibid.*

We might think we are in the same semantic waters with the intuitions that our experiment has uncovered. For our intuitions about moral responsibility seem to vary with the actuality of the world we are considering. On a two-dimensional interpretation, the results may be explained as follows: On the assumption that our world is indeterministic, then moral responsibility requires indeterminism in other worlds, but if our world is in fact deterministic, then moral responsibility does not require indeterminism. A conditional analysis of the concept of moral responsibility would then look like this:

> If the actual world is indeterministic, then moral responsibility is incompatible with determinism; else in the actual world compatibilism is true and is true of moral responsibility in other worlds.

That is, the intuition that we are in fact morally responsible is a nonnegotiable intuition. Our concept of moral responsibility, and our judgments about responsibility in other worlds rides upon what in fact holds in the actual world.

The advantage of the analytical conditional analysis is that it provides us with an account of our concept that (1) seems to respect the pattern of intuitions that the folk have, and (2) makes it the case that the correct analysis of moral responsibility depends upon whether or not determinism is true in our world, but makes it possible to grasp the concept in the absence of information as to whether or not that is the case.

On a 2D approach to our results, one might maintain that the A-intension for the proposition that people are morally responsible is never rendered false by the status of determinism when evaluated in a world taken as actual; that is, regardless of whether the world is determinist or indeterminist, the A-intension yields compatibilist judgments.[41] C-intensions need not be true: if the world taken as actual is not deterministic, then in counterfactual deterministic worlds agents may not be judged to be free or held morally responsible.

Since the 2-D approach in analytic metaphysics purports to give an account of the folk concept, this view generates some predictions. First, in our experiments, subjects were only asked about their intuitions about moral responsibility in one universe or another. The prediction that results is that if

---

[41] Of course, our data do not show anything approaching complete unanimity in intuitions, and Derk Pereboom has suggested (personal communication) that this raises important questions about the status of the 2D explanation of our findings. On the 2D account, how do we evaluate the minority of responses in the *Actual* condition that are incompatibilist? Are such judgments mistaken? This implicates delicate issues about how to interpret individual differences. Ideally, one would like to know whether the variation tracks stable features of the individual, or whether the variation is just noise. In the case of some philosophically-relevant intuitions, the individual differences do seem to be stable. See S. Nichols and J. Ulatowski, J. "Intuitions and Individual Differences: The Knobe Effect Revisited" *Mind and Language*, XXII, (2007) p. 346-365. But in the present case, we don't have enough information to know whether the individual differences in intuitions about moral responsibility are stable or random, and so we are unable to begin to evaluate whether the minority respondents should be regarded as mistaken in their responses.

people were asked to make judgments about two cases in succession, with one scenario presented as the actual world, their judgments about the responsibility of agents in the second case would be modulated by their initial judgments about the description they are given of the actual world. That is, if, in the actual world, scientists discovered that determinism were true, subsequent judgments about other deterministic worlds would yield judgments assigning high agreement to assertions of moral responsibility, whereas if scientists determined that the actual world was in fact an indeterministic world, then judgments about other universes that were deterministic would show the pattern we found. Presumably the same sort of manipulation could be performed by stipulating that agents in this world or other worlds do or do not have free will, in order to examine the dependence of moral responsibility on freedom itself, as opposed to determinism and indeterminism.

Wider implications for thought experiments

Our experimental manipulation has demonstrated that people's intuitions about moral responsibility vary across universes. Taken as some brute fact, this pattern of responses may be interpreted to suggest that judgments are subject to framing effects that affect outcomes in surprising and somewhat random ways. If so, a natural conclusion to reach is that intuitions are fragile, corruptible, and inconsistent, and that perhaps probing those intuitions is not the best guide to understanding philosophically important concepts such as freedom or moral responsibility. The methodological upshot of such a conclusion might to be to reject folk intuitions wholesale, and to concentrate on pushing for a coherent analysis of a concept that holds across worlds, regardless of its coherence with folk judgments. The fact that these judgments vary unexpectedly across worlds may also be thought to call into question traditional philosophical reliance on possible worlds analyses for understanding our concepts.

However, we think that such a conclusion too hastily dismisses the value of probing intuitions. The emotion-based model from psychology and neuroscience can causally explain the pattern of results we get. If this explanation is correct, we may expect to see a divergence in patterns of metaphysical judgments across worlds. In affect-neutral cases, such as for instance many cases involving mereology, constitution, or natural kinds, the affect-based explanation would predict no pattern of deviation across worlds, but in cases where affect is likely to be elicited, such as, for instance, many cases involving freedom, personal identity, moral propriety, or the existence of God, one might expect to see a pattern like the one we found here. Whether this is the case is an empirically tractable question: people's judgments about such cases across worlds can be investigated, and in order to more explicitly test the affect hypothesis, these results can be correlated with resultant levels of affect or the activity of emotion-related brain areas.[42]

[42] We do see a varied pattern of judgments about natural kinds across worlds (see, for example, Kripke's *Naming and necessity*)Cambridge MA: Harvard University Press, 1972) and Putnam's "The Meaning of 'Meaning"' In *Mind, Language and Reality* (Cambridge: Cambridge University Press, 1975) p. 215-271. However, there is no obvious reason to expect these judgments to be driven by emotional responses to the possible worlds scenarios. This suggests that the emotion-based model is not sufficient for explaining the pattern of judgments for all our modal intuitions. There may be other factors that also lead to differences in cross-world judgments.

It is not only empirical methods that provide greater insight into our results. Careful thought reveals that the pattern of judgments we in fact see when doing experimental philosophy is not merely random and incoherent, but itself tells us something interesting: our judgments about moral responsibility are dependent not only on the nature of the world of the agent about which we are making a judgment, but also on the nature of our own world. This realization may in fact be an empirical advertisement for 2-D possible worlds semantics, and may call for taking the two-dimensionalist position seriously.

If the 2-D model is taken to be an analysis of our concept of moral responsibility, as 2-D interpretations often are, then it seems that it would preclude the affect explanation of the results. On the other hand, the affect explanation has independent support from experiments from Nichols and Knobe. It is still early days in the exploration of folk intuitions, and we are reluctant to rule out models prematurely. One possibility might be to take the analytical conditional analysis to be a systematization of people's judgments under various conditions, and to turn to the affect model for an explanation of the genesis of that pattern.

Finally, note that if either the affect-based model or the analytic conditional analysis is correct, then by understanding folk judgments we learn something interesting and important about the nature of our concept of moral responsibility. We also come to understand something about the genesis of one of the perennial debates in philosophy.

Final thoughts

We began by considering folk intuitions about the relation between moral responsibility and determinism. Are people compatibilists or incompatibilists? If the affect model suggested by neuroscience and psychology is correct, it seems plausible to deny that people are really compatibilists, for they are inclined to disagree with the claim that people are morally responsible in deterministic worlds considered as counterfactual. The finding that subjects agree that people are morally responsible (and, to a lesser extent free) in a deterministic universe if that universe is considered as actual may be explained by the biasing effect of affect, which is disproportionately recruited in assessing questions made relevant for the subject. On the 2D model, however, we might plausibly maintain that people are intuitive compatibilists, if we think that questions about freedom and responsibility are meant to be assessed in a world taken as actual, for on the 2D reading, in the world taken as actual people are morally responsible regardless of the truth of determinism: the A-intension is always compatibilist. However, the 2D model might better be taken to suggest that both the traditional compatibilist and incompatibilist philosophical positions are too narrow to properly capture the subtlety of our intuitions, and that a 2D semantics, or something like it, is required to adequately capture the intuitions of the folk. This would suggest, moreover, that there is something very natural about a 2D reading, and that people need not be trained in philosophical technicalia in order to embody both actual and counterfactual possibilities in their untutored concepts.

As for responsibility, our results indicate that should neuroscience or philosophy lead the folk to come to think, correctly or mistakenly, that our minds are mechanistic and our choices are

determined, our judgments about moral responsibility will remain largely intact. We should not be deterred from a scientific appreciation of the mind by fears of nihilism or social disintegration.

Adina L. Roskies
Department of Philosophy
Dartmouth College

Shaun Nichols
Department of Philosophy
University of Arizona

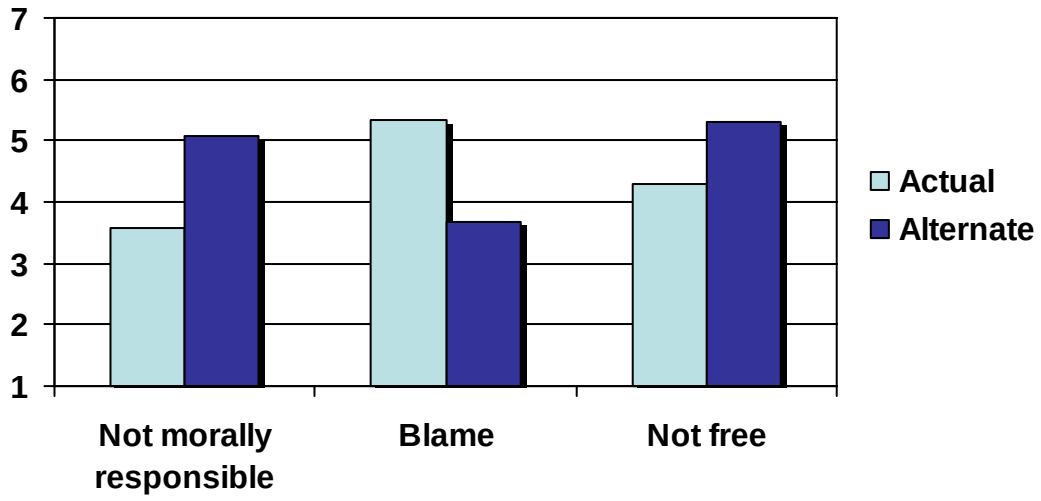

Figure legend:

This graph shows average ratings for level of agreement with statements about the moral responsibility, blameworthiness, and freedom of agents in deterministic worlds if those worlds are the actual world or a world in some alternate universe. Ratings of 1 correspond to *disagree completely*, 7 with *agree completely*, and 4 is neutral. Level of agreement was assessed to the following questions: in such a world (1) it is impossible for a person to be fully morally responsible for their actions; (2) people should still be morally blamed for committing crimes; and (3) it is impossible for people to make truly free choices.