Compatibilism Can Be Natural

John Turri

Philosophy Department and Cognitive Science Program
University of Waterloo
200 University Avenue West
Waterloo, Ontario N2L3G1, Canada
john.turri@gmail.com

Abstract: Compatibilism is the view that moral responsibility is compatible with determinism. Natural compatibilism is the view that in ordinary social cognition, people are compatibilists. Researchers have recently debated whether natural compatibilism is true. This paper presents six experiments (N = 909) that advance this debate. The results provide the best evidence to date for natural compatibilism, avoiding the main methodological problems faced by previous work supporting the view. In response to simple scenarios about familiar activities, people judged that agents had *moral responsibilities to perform actions* that they were unable to perform (Experiment 1), were *morally responsible for* unavoidable outcomes (Experiment 2), were *to blame for* unavoidable outcomes (Experiments 3-4), *deserved blame for* unavoidable outcomes (Experiment 5), and *should suffer consequences for* unavoidable outcomes (Experiment 6). These findings advance our understanding of moral psychology and philosophical debates that depend partly on patterns in commonsense morality.

Keywords: determinism; responsibility; agency; moral psychology; folk metaphysics

Introduction

Attitudes toward freedom and moral responsibility have important social consequences (Monroe, Dillon & Malle 2014). For instance, they can influence people's job performance, academic performance, and frequency of prosocial behavior (Stillman, Baumeister, Vohs, Lambert, Fincham & Brewer 2010; Feldman, Chandrashekar & Wong 2016; Baumeister, Masicampo & DeWall 2009). In light of their social importance, it is no surprise that these issues have been studied extensively in the humanities and social sciences. For example, philosophers and psychologists have long debated the merits of compatibilism and its denial, incompatibilism. Compatibilism is the view that acting freely and moral responsibility are compatible with determinism.

The theoretical debates surrounding these issues can become extremely complicated, often turning on subtle matters concerning topics ranging from cognitive neuroscience to quantified modal logic (for reviews, see McKenna 2009; O'Connor 2010; Vihvelin 2011). Nevertheless, one aspect of the debate has remained firmly rooted in commonsense, and it is this aspect of the debate that I focus on here. It is often claimed that compatibilism or incompatibilism is a natural part of ordinary social cognition (e.g. Hume 1748/1993; Reid 1785; Kane 1999, p. 217; Pereboom 2001, p. xvi; Roskies & Nichols 2008; Rose & Nichols 2013). That is, it is often claimed that our commonsense moral psychology is implicitly committed to one view or the other.

The question about commonsense moral psychology is important for at least two reasons. On the one hand, understanding moral psychology is an important part of understanding human psychology overall. Indeed, it might be argued that moral psychology is one of the more fascinat-

ing aspects of human mentality and culture, because it is so unlike anything else observed in the animal world, even if it certainly has origins in more primitive instincts and mechanisms shared with other primates and mammals more generally (Haidt 2007; de Waal 2006). On the other hand, the theoretical debate has often assumed that the burden of proof rests with the side that contradicts commonsense (for a review, see Nahmias, Morris, Nadelhoffer & Turner 2006). For example, if ordinary moral psychology assumes that compatibilism is true, then incompatibilists will need stronger arguments to persuade us that their position is correct.

Several recent studies have begun examining the status of compatibilism and incompatibilism in commonsense moral psychology (e.g. Nahmias, Morris, Nadelhoffer & Turner 2005; Woolfolk, Doris & Darley 2006; Nichols & Knobe 2007; Sarkissian et al. 2010; Cova & Kitano 2014; for reviews, see Sommers 2010 and Nichols 2011). The results have been mixed, with some suggesting that people are natural compatibilists and some suggesting that they are natural incompatibilists (for a review, see Deery, Davis & Carey 2014; see also Schulz, Cokely & Feltz 2011; May 2014).

However, prior work supporting natural compatibilism suffers from several methodological concerns. First, it uses long, complicated, and incredible stimuli. Second, as has been previously noted (Nichols & Knobe 2007), some stimuli are provocative, raising the worry that results are due to emotional interference and performance error (e.g. one case involves egregious marital infidelity, terrorists hijacking a plane at gunpoint, the execution of an elderly person, and an agent who "blow[s] his friend's brains out" with a pistol; see Woolfolk, Doris & Darley 2006). Third, control conditions were only loosely matched (e.g. one manipulation consisted of 15

words being exchanged for 46 words; see Nahmias, Shepard & Reuter 2014, Appendix, scenario for Experiment 1, penultimate paragraph). Fourth, the studies did not assess whether participants understood variables in the relevant way. This is important because recent work suggests that our naive understanding of psychological and physical processes is surprisingly indeterministic. In one recent study, when asked to assess the probability of an "inevitable" and "causally determined" outcome, people rated it between 70% and 85% likely (Turri in press; see also Nichols 2004; Rose & Nichols 2013).

If people's interpretation of scenarios is often surprisingly indeterministic, then it complicates attempts to assess whether people are natural compatibilists or incompatibilists. Indeed, it suggests another possibility: pre-theoretically, the question of compatibilism might be irrelevant and thus never arise. If indeterminism is assumed to be true, then it would not matter whether responsibility, or anything else, is compatible with determinism. Simply put, ordinary social cognition might never confront the question. To circumvent this worry, it is important gather people's judgments about moral responsibility in contexts where they agree that the agent cannot perform the relevant action.

Motivated by that possibility, the goal of the present research is to gain better evidence about natural compatibilism regarding moral responsibility. In order to avoid the methodological concerns raised above, I conducted six experiments using simple, clear, short, and closely matched stimuli about familiar actions. I also included multiple measures to assess how people understood key variables. The results provide the best evidence to date for natural compatibilism. More specifically, the results provide evidence for natural compatibilism about five categories

connected to moral responsibility, including having a moral responsibility to perform an action (Experiment 1), being responsible for an outcome (Experiment 2), being to blame for an outcome (Experiments 3 and 4), deserving blame for an outcome (Experiment 5), and deserving to suffer (Experiment 6). I do not assume that these five exhaust the list of potentially relevant categories. I chose them because they are common moral judgments that, I think, are intrinsically interesting, and because they are implicated in the theoretical literature on "determinism and moral responsibility" (e.g., see Vihvelin 2011; Russell 2014). Studying them all in the same context provides an opportunity to discern whether, say, natural compatibilism is true for some but not all of them, or whether it better captures the central tendencies for some.

Each experiment tests the principal predictions of natural incompatibilism and natural compatibilism for a specific moral status. In each case, natural incompatibilism predicts that people will deny that agents have the relevant status, whereas natural compatibilism predicts that people will attribute such a status. For example, natural incompatibilism about *having a moral responsibility* predicts that when it is impossible for an agent to perform an action, people will deny that the agent has a responsibility to perform it; by contrast, natural compatibilism predicts that people will agree that the agent has a responsibility. Of course, the predictions of natural incompatibilism depend upon people understanding the cases in the relevant way. For instance, if people reject the assumption that the agent cannot perform the relevant action, then natural incompatibilism does not predict that people will deny moral responsibility. The point can also be expressed in a way that respects the fact that these judgments come in degrees. Natural incompatibilism predicts that people's judgments about ability will strongly constrain their attribution of

the relevant moral status. More specifically, the prediction is that attributions of the relevant moral status will not significantly exceed attributions of the ability to perform the relevant action.

Experiment 1

Method

Participants

Two hundred and three participants were tested (aged 18-72 years, mean age = 34 years; 96 female; 93% reporting English as a native language). Participants were U.S. residents, recruited and tested online using Amazon Mechanical Turk (AMT) and Qualtrics, and compensated $0.40 for approximately 2 minutes of their time. The same recruitment and compensation procedures were used for all experiments reported in this paper. Repeat participation was prevented within and across experiments (by AMT worker ID).

Materials and Procedure

Participants were randomly assigned to one of four conditions in a 2 (action type: delivery, evaluation) × 2 (status: impossible, guaranteed) between-subjects design. All participants read a simple story, responded to three test items, then completed a brief demographic questionnaire.

The status factor manipulated whether it was impossible or guaranteed that the agent could

perform the action. The action type factor manipulated which action the story focused on. I had no expectations for this factor and included it as a robustness check. The story for the delivery conditions concerned a man who promised to deliver a package. The story for the evaluation conditions concerned a woman evaluating an employee. Here is the text of the stories (status manipulation in brackets):

> (Delivery) A man promised to deliver a package by 4pm. He just got on the freeway. Given current traffic conditions, it is physically [impossible/guaranteed] that he will deliver the package by 4pm. As a matter of physics, it is literally [impossible/guaranteed] that he can make it by 4pm.
>
> (Evaluation) A woman is evaluating her employee's performance. The employee performed excellently. Given the current condition of the woman's brain, it is physically [impossible/guaranteed] that she will give the employee a positive evaluation. As a matter of brain chemistry, it is literally [impossible/guaranteed] that she can give the employee a positive evaluation.

I used the evaluation story it because previous research has shown its status manipulation to be effective (Turri 2016), which is critical to answering the present research question (see the Introduction). In particular, it is essential that participants accept that the agent cannot perform the action described as impossible. The status manipulation for the delivery story was validated in a pilot study.

After reading the story, participants responded to three test items while the story remained at the top of the screen. They first responded to two statements in a matrix table, the order of

which was randomized:

1. The man could still deliver the package by 4pm. / The woman could still give the employee a positive evaluation. (could)
2. The man has a moral responsibility to deliver the package by 4pm. / The woman has a moral responsibility to give the employee a positive evaluation. (responsibility)

Response to these items was collected on a standard 7-point Likert scale, 1 ("strongly disagree") – 7 ("strongly agree"), left-to-right across the participant's screen. Participants then proceeded to a new screen and completed a percentage task:

3. On a scale of 0% to 100%, how likely is it that [the man will deliver the package by 4pm / the woman will give the employee a positive evaluation]? (percent)

Response to this item was collected in a text box directly below the question. After testing, participants completed a brief demographic questionnaire.

Results

Preliminary analyses of variance treating action type as a random factor revealed no effect of action type on the dependent measures, ps > .656. Accordingly, because action type was included merely as a robustness check and is not of independent theoretical interest, the analyses that follow collapse across that factor. The status manipulation (impossible/guaranteed) was extremely effective. (See Figure 1 and Table 1.) Participants denied that the agent could perform the relevant action in the impossible condition, and they agreed in the guaranteed condition. (See Table 2.) Mean response to the responsibility statement was above the midpoint in both the impossible

and possible conditions. Paired samples t-tests showed that in the impossible condition, mean response to the responsibility statement exceeded mean response to the could statement; by contrast, in the guaranteed condition, mean response to the could statement exceeded mean response to the responsibility statement. (See Table 3.)

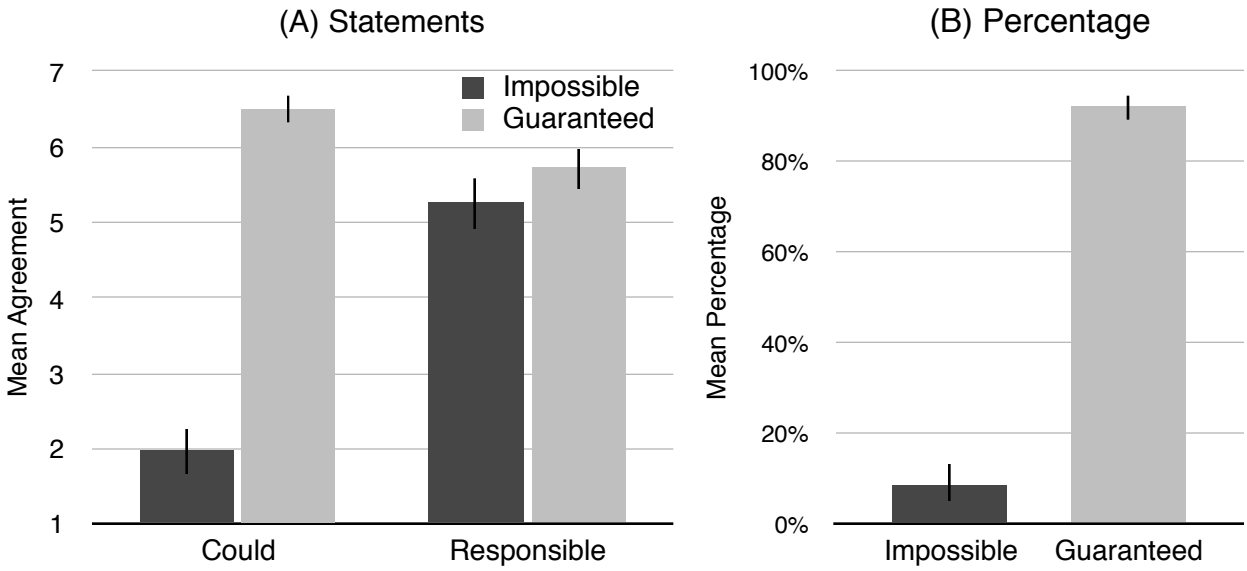

Figure 1. Experiment 1. Panel A: mean response to the test statements about whether the agent could still perform the action, and whether the agent has a moral responsibility to perform the action; the scale ran 1 (SD) - 7 (SA). Panel B: mean estimate of the percentage chance that the agent will perform the action; the scale ran 0-100%. Error bars show 95% confidence intervals. Results collapse across action type (delivery/evaluation).

Table 1. Experiment 1. Independent samples t-tests comparing responses to the three dependent measures in the impossible and guaranteed conditions.

	Impossible		Guaranteed							
Measure	**M**	**SD**	**M**	**SD**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	1.96	1.52	6.51	0.98	-25.36	172.31	<.001	-4.55	-4.91, -4.20	3.86
responsible	5.25	1.66	5.71	1.42	-2.16	201	.032	-0.47	-0.90, -0.40	0.30
percentage	8.81	21.54	92.34	14.15	-32.68	174.66	<.001	-83.53	-88.57, -78.48	4.94

Table 2. Experiment 1. One sample t-tests on mean response to the could and responsibility statements in the impossible and guaranteed conditions. Test value = 4.

	Impossible						Guaranteed					
Measure	**t**	**df**	**p**	**MD**	**95% CI**	**d**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	-13.54	101	<.001	-2.04	-2.34, -1.74	1.34	25.90	100	<.001	2.52	2.32, 2.71	2.57
responsible	7.60	101	<.001	1.25	0.92, 1.57	0.75	12.09	100	<.001	1.71	1.43, 1.99	1.20

Table 3. Experiment 1. Paired samples t-tests comparing mean response to the could and responsibility statements in the impossible and guaranteed conditions.

Condition	**t**	**df**	**p**	**MD**	**95% CI**	**d**
impossible	-14.78	101	<.001	-3.28	-3.73, -2.84	1.47
guaranteed	5.97	100	<.001	0.80	0.54, 1.07	0.62

I also analyzed the data from only those participants in the impossible condition who answered “strongly disagree” to the could statement and “0%” on the percentage task. In this group (N = 51), mean agreement was above the midpoint for the responsibility statement (M = 5.25, SD = 2.03), $t(50) = 4.42$, $p < .001$, MD = 1.26, d = 0.62.

Finally, I compared response to the responsibility statement from two groups. One group included those in the impossible condition who answered “strongly disagree” to the could statement and “0%” on the percentage task (N = 51) (i.e. the same group described in the previous paragraph). The other group included those in the guaranteed condition who answered “strongly agree” to the could statement and “100%” on the percentage task (N = 41). Mean response was lower for the first group (M = 5.25/6.24, SD = 2.03/1.02), independent samples t-test, $t(76.87) = -3.04$, $p = .003$, MD = -0.99, d = 0.65. Nevertheless, for both groups, mean response was signifi-

cantly above the midpoint, one sample t-tests, ps < .001, and modal response was "strongly agree."

Discussion

People agreed that an agent who could not perform an action was still responsible for performing it. This rate of responsibility attribution was slightly lower than but still comparable to the rate observed in closely matched control conditions where people overwhelmingly agreed that the agent could perform the action. The same basic pattern occurred for two very different actions. These results support the conclusion that when judging responsibility in a range of ordinary cases, people are natural compatibilists.

The next experiment investigates whether a similar pattern occurs if participants are asked to judge whether agents are *responsible for* their actions, as opposed to having a responsibility to perform the actions.

Experiment 2

Method

Participants

Two hundred new participants were tested (aged 19-69 years, mean age = 35 years; 97 female; 97% reporting English as a native language).

Materials and Procedure

Participants were randomly assigned to one of four conditions in the same 2 (action type: delivery, evaluation) × 2 (status: impossible, guaranteed) between-subjects design from Experiment 1. The procedures were exactly the same as in Experiment 1. The materials differed in three ways. First, in the story for delivery conditions, instead of reading "it is physically [impossible/guaranteed] that he *will* deliver the package by 4pm," it read "it is physically [impossible/guaranteed] that he *can* deliver the package by 4pm" (participants did not see the italics, which are just for illustration); a comparable switch from "will" to "can" was made to the story for the evaluation conditions. The purpose of this switch was to consistently allow for the next change. Second, at the end of the story for delivery conditions, an additional sentence appeared, "He will arrive late"; a comparable sentence was added to the end of the story for evaluation conditions, "She will give the employee a negative evaluation." The purpose of this change was to allow participants to rate the agent's moral responsibility for a definite action indicated in the scenario. Third, the responsibility statement was different:

> The man is morally responsible for the time he delivers the package. / The woman is morally responsible for the evaluation she gives the employee.

The could statement and percentage task were the same as in Experiment 1.

Results

As in Experiment 1, preliminary analyses of variance treating action type as a random factor revealed no effect of action type on the dependent measures, ps > .250. Accordingly, because ac-

tion type was included merely as a robustness check and is not of independent theoretical interest, the analyses that follow again collapse across that factor. The status manipulation was extremely effective. (See Figure 2 and Table 5.) Participants denied that the agent could perform the relevant action in the impossible condition, and they agreed in the guaranteed condition. (See Table 5.) Mean response to the responsibility statement was above the midpoint in both conditions. Paired samples t-tests showed that in the impossible condition, mean response to the responsibility statement exceeded mean response to the could statement; by contrast, in the guaranteed condition, mean response to the two statements did not differ. (See Table 6.)

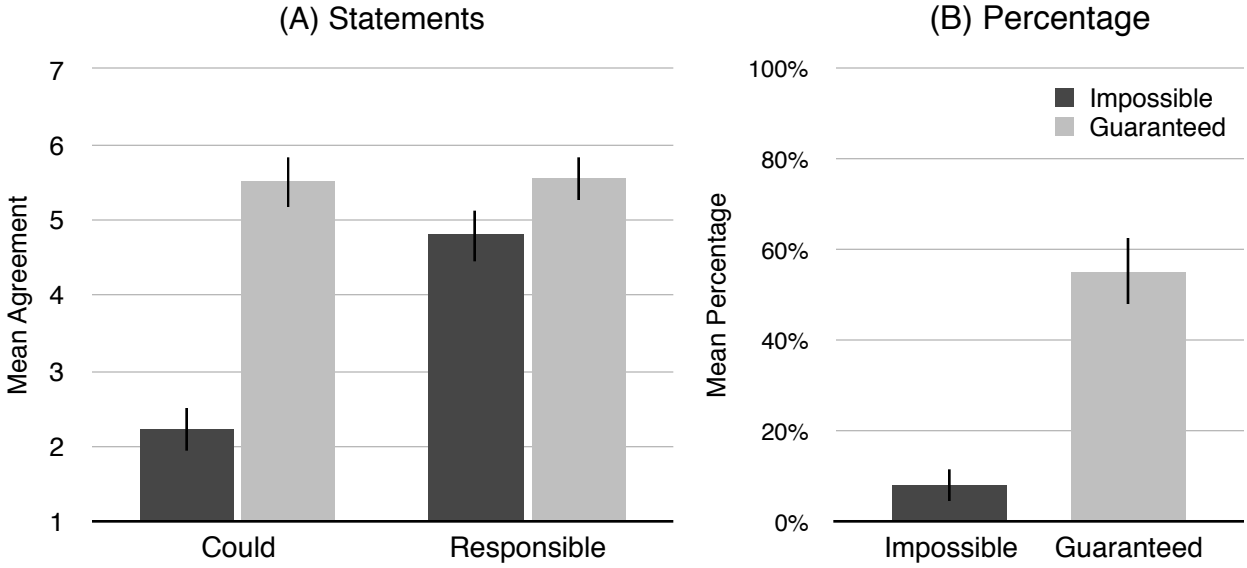

Figure 2. Experiment 2. Panel A: mean response to the test statements about whether the agent could still perform the action, and whether the agent is responsible for the outcome; the scale ran 1 (SD) - 7 (SA). Panel B: mean estimate of the percentage chance that the agent will perform the action; the scale ran 0-100%. Error bars show 95% confidence intervals. Results collapse across action type (delivery/evaluation).

Table 4. Experiment 2. Independent samples t-tests.

	Impossible		Guaranteed							
Measure	**M**	**SD**	**M**	**SD**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	2.21	1.97	5.51	1.64	-12.90	198	<.001	-3.30	-3.80, -2.80	1.84
responsible	4.80	1.82	5.56	1.39	-3.32	185	.001	-0.76	-1.21, -0.31	0.49
percentage	7.75	18.64	55.30	37.51	-11.35	145.1	<.001	-47.55	-55.83, -39.27	1.88

Table 5. Experiment 2. One sample t-tests. Test value = 4.

	Impossible						Guaranteed					
Measure	**t**	**df**	**p**	**MD**	**95% CI**	**d**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	-9.11	99	<.001	-1.79	-2.18, -1.40	0.91	9.23	99	<.001	1.51	1.19, 1.83	0.92
responsible	4.40	99	<.001	0.80	0.44, 1.16	0.44	11.24	99	<.001	1.56	1.28, 1.84	1.12

Table 6. Experiment 2. Paired samples t-tests comparing mean response to the responsibility and could statements in the four conditions.

Condition	**t**	**df**	**p**	**MD**	**95% CI**	**d**
impossible	11.99	99	<.001	2.59	2.16, 3.02	1.20
guaranteed	0.26	99	.796	0.05	-0.33, 0.43	0.03

I also analyzed the data from only those participants in the impossible condition who answered “strongly disagree” to the could statement and “0%” on the percentage task. In this group (N = 51), mean agreement was above the midpoint for the responsibility statement (M = 4.59, SD = 1.82), t(50) = 2.30, p = .025, MD = 0.59, d = 0.34. Modal response was “slightly agree.”

Unlike in Experiment 1, I could not conduct a meaningful comparison between the group just discussed and those in the guaranteed condition who answered “strongly agree” to the could statement and “100%” on the percentage task. This was because only 8 participants in the guar-

anteed condition met these criteria. This difference is due to the switch from "will" to "can" described in the Method section for the present experiment. The same is true for the guaranteed conditions in all subsequent experiments reported here.

Discussion

People agreed that an agent who could not have avoided an outcome was still morally responsible for it. People more strongly agreed with the responsibility attribution when the agent could have avoided the outcome, which suggests that information about ability informs responsibility attributions, though not nearly to the extent one would expect if people were natural incompatibilists in this regard. Overall, the results from this experiment provide evidence for natural compatibilism about being morally responsible for unavoidable outcomes.

The next experiment investigates whether a similar pattern occurs if participants are asked to judge whether agents are *to blame* for their actions.

Experiment 3

Method

Participants

Two hundred and two participants were recruited and tested (aged 18-75 years, mean age = 35 years; 110 female; 93% reporting English as a native language).

Materials and Procedure

All procedures and materials were exactly the same as in Experiment 2, except that the responsibility statement was replaced with a blame statement:

> The man is to blame for the time he delivers the package. / The woman is to blame for the evaluation she gives the employee.

Results

As in Experiments 1 and 2, preliminary analyses of variance treating action type as a random factor revealed no effect of action type on the dependent measures, ps > .252. Accordingly, because action type was included merely as a robustness check and is not of independent theoretical interest, the analyses that follow again collapse across that factor. The status manipulation was extremely effective. (See Figure 3 and Table 7.) Participants denied that the agent could perform the relevant action in the impossible condition, and they agreed in the guaranteed condition. (See Table 8.) Mean response to the blame statement was above the midpoint in the guaranteed condition and no different from the midpoint in the impossible condition. Paired samples t-tests showed that in the impossible condition, mean response to the blame statement exceeded mean response to the could statement; by contrast, in the guaranteed condition, mean response to the two statements did not differ. (See Table 9.)

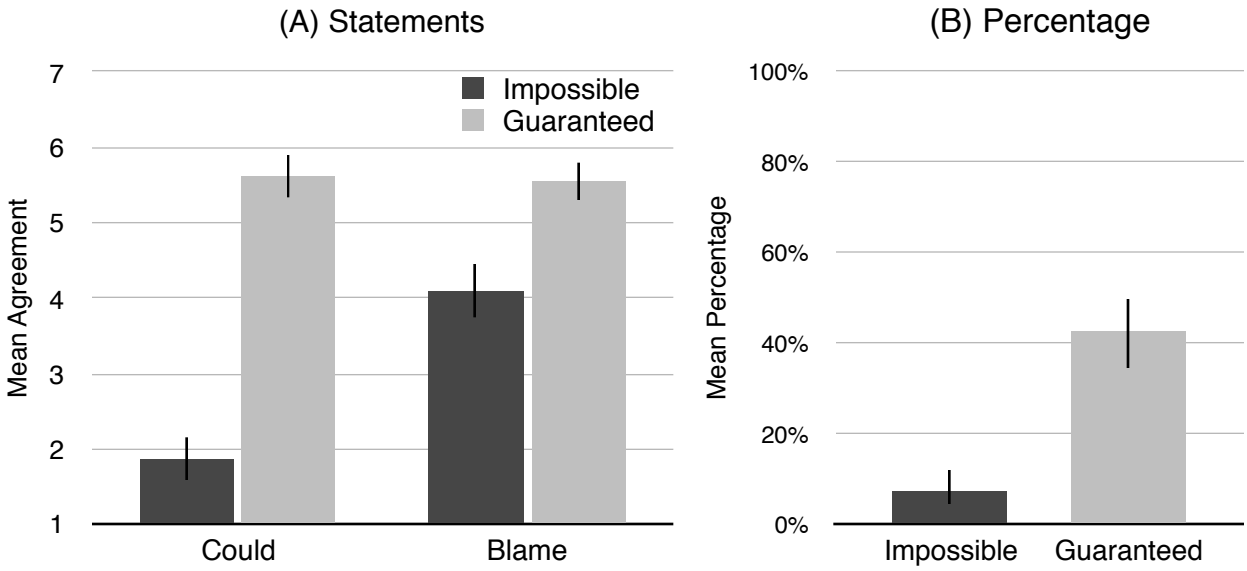

Figure 3. Experiment 3. Panel A: mean response to the test statements about whether the agent could avoid the outcome, and whether the agent was to blame; the scale ran 1 (SD) - 7 (SA). Panel B: mean estimate of the percentage chance that the agent will perform the action; the scale ran 0-100%. Error bars show 95% confidence intervals. Results collapse across action type (delivery/evaluation).

Table 7. Experiment 3. Independent samples t-tests comparing responses to the three dependent measures in the impossible and guaranteed conditions.

	Impossible		Guaranteed							
Measure	**M**	**SD**	**M**	**SD**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	1.87	1.36	5.61	1.41	-19.21	200	<.001	-3.74	-4.13, -3.36	2.72
blame	4.11	1.87	5.55	1.54	-6.00	192.87	<.001	-1.44	-1.92, -0.97	0.86
percentage	7.64	19.49	42.66	37.39	-8.35	150.59	<.001	-34.02	-43.31, -26.73	1.36

Table 8. Experiment 3. One sample t-tests on mean response to the could and blame statements in the impossible and guaranteed conditions. Test value = 4.

	Impossible						Guaranteed					
Measure	**t**	**df**	**p**	**MD**	**95% CI**	**d**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	-15.72	100	<.001	-2.13	-2.40, -1.86	1.57	11.53	100	<.001	1.61	1.34, 1.89	1.14
responsible	0.59	100	.560	0.11	-0.26, 0.48	0.06	10.15	100	<.001	1.55	1.25, 1.86	1.01

Table 9. Experiment 3. Paired samples t-tests comparing mean response to the could and blame statements in the impossible and guaranteed conditions.

Condition	t	df	p	MD	95% CI	d
impossible	-10.84	100	<.001	-2.24	-2.65, -1.83	1.10
guaranteed	0.34	100	.736	0.06	-0.29, 0.41	0.03

I also analyzed the data from only those participants in the impossible condition who answered "strongly disagree" to the could statement and "0%" on the percentage task. In this group (N = 54), mean agreement did not differ from the midpoint for the responsibility statement (M = 3.78, SD = 2.13), t(53) = -0.77, p = .446, n.s.

Discussion

People were neutral on whether an agent was to blame for an unavoidable outcome. The rate of blame attribution was statistically lower than that observed in a closely matched control condition where the outcome was avoidable. These results do not support natural compatibilism about blame. At the same time, the results do not support natural incompatibilism about blame either. That is, when the outcome could not be avoided, the central tendency was not to deny that the agent was to blame. Instead, the central tendency was ambivalence about blame.

A possible explanation of this finding, consistent with natural compatibilism about blame, is that people needed more information about the agent or situation before expressing a judgment about blame. Previous work has shown that whether an agent "identifies" with an outcome affects people's willingness to attribute responsibility (Woolfolk, Doris & Darley 2006; Pizarro,

Uhlmann & Salovey 2003; see also Frankfurt 1998, Watson 1996). Identifying with an outcome involves wanting it to happen or endorsing it upon reflection. Thus it is possible that people in this experiment were ambivalent about blame because they wanted to know more about how the agent viewed the outcome. If the agent would endorse the unavoidable outcome upon reflection, then perhaps people will attribute blame. The next experiment tests this possibility. More specifically, it tests natural compatibilism about blame in contexts where the agent endorses an unavoidable outcome.

Because action type did not matter in the first three experiments and is not of independent theoretical interest, the remaining experiments focus on a single action (i.e. do not include action type as an independent variable).

Experiment 4

Method

Participants

One hundred participants were recruited and tested (aged 20-68 years, mean age = 34 years; 59 female; 90% reporting English as a native language).

Materials and Procedure

Participants were randomly assigned to one of two conditions (rejection, identification) in a between-subjects design. The procedures were exactly the same as in Experiment 3. The materials

were the same as those for the impossible evaluation condition in Experiment 3, except that a subordinate clause was added to the final sentence. Here is the final sentence for the current stimuli (rejection/identification manipulation in brackets):

> She will give the employee a negative evaluation, which is something that, upon reflection, she would [not/fully] endorse doing.

Participants then responded to the same three dependent measures as in previous experiments (could statement, blame statement, percentage task).

Results

Independent samples t-tests revealed no effect of condition on any of the three dependent measures. (See Table 10.) Accordingly, the following analyses collapse across condition. Mean response to the could statement was below the midpoint, whereas mean response to the blame statement was above the midpoint. (See Figure 4 and Table 11.) A paired samples t-test showed that mean response to the blame statement ($M = 4.78$, $SD = 2.01$) exceeded mean response to the could statement ($M = 3.06$, $SD = 2.13$), $t(99) = 6.50$, $p < .001$, $MD = 1.72$, $d = 0.65$.

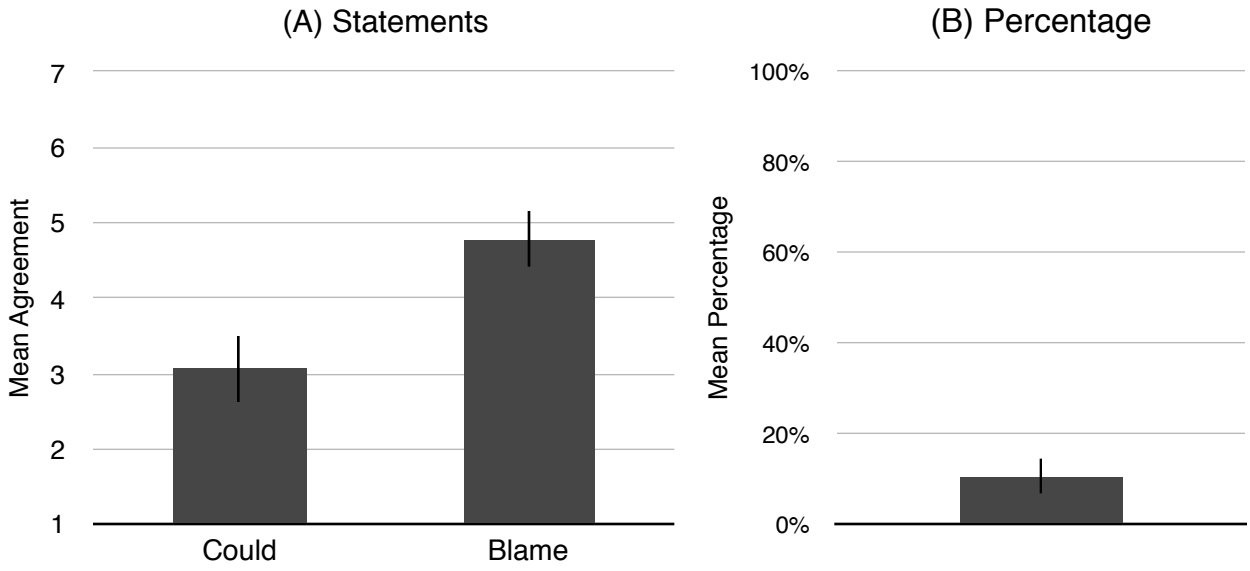

Figure 4. Experiment 4. Panel A: mean response to the test statements about whether the agent could still avoid the outcome, and whether the agent was to blame; the scale ran 1 (SD) - 7 (SA). Panel B: mean estimate of the percentage chance that the agent will avoid the outcome; the scale ran 0-100%. Error bars show 95% confidence intervals. Results collapse across condition (rejection/identification).

Table 10. Experiment 4. Independent samples t-tests comparing responses to the three dependent measures in the rejection and identification conditions.

	Rejection		Identification							
Measure	**M**	**SD**	**M**	**SD**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	3.00	2.12	3.12	2.16	-0.28	98	.784	-0.12	-0.97, 0.73	0.06
blame	4.73	2.01	4.82	2.04	-0.22	98	.827	-0.09	-0.89, 0.71	0.04
percentage	9.08	19.64	11.84	22.88	-0.65	98	.520	-2.76	-11.24, 5.72	0.13

Table 11. Experiment 4. One sample t-tests on mean response to the could and blame statements. Test value = 4.

Measure	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	-4.41	99	<.001	-0.94	-1.36, -0.52	0.44
blame	3.88	99	<.001	0.78	0.38, 1.18	0.39

I also analyzed the data from only those participants who answered "strongly disagree" to the could statement and "0%" on the percentage task. In this group (N = 27), mean agreement with the blame statement (M = 4.93, SD = 2.35) was higher than in the sample as a whole and above the midpoint, $t(26) = 2.05$, $p = .051$, MD = 0.93. Modal response was "strongly agree."

Discussion

People agreed that an agent who could not avoid an outcome was still to blame for it. This result supports the conclusion that when judging blame in some ordinary cases, people are natural compatibilists. More specifically, it supports this conclusion when people are given information about how the agent herself views the outcome. This contrasts with the findings from Experiment 3, where people were not given information about how the agent viewed the outcome, and they were neutral on whether the agent was to blame.

An unexpected result from this experiment is that the difference between an agent rejecting and endorsing the outcome upon reflection did not affect blame attributions. This contrasts with previous work that did find a difference (Woolfolk, Doris & Darley 2006). There could be many explanations for this divergence in findings, including differences in stimulus length and selection selection. Another possibility is that participants in the present experiment attributed blame in the two conditions for different reasons. On the one hand, when the agent would endorse the outcome upon reflection, this, in line with previous findings, increased participants' confidence that the agent is to blame. By contrast, when the agent would reject the outcome upon reflection,

participants took this as evidence that the agent thought she herself was to blame, which then increased participants' confidence that she was to blame. Yet another possibility is that simply adding more detail about the agent's psychology increases people's willingness to blame. This is consistent with earlier work reporting that more concrete stimuli elicit more compatibilist responses (Nichols & Knobe 2007), because the added psychological details plausibly make the case more concrete. Further work is required to understand what relationship, if any, exists between the present findings and earlier results. Whatever the ultimate explanation, for present purposes, the important point is that adding either of two small pieces of information to the case changed the central tendency from ambivalence (Experiment 3) to agreement.

The next experiment investigates whether a similar pattern occurs if participants are asked to judge whether agents *deserve blame* for their actions.

Experiment 5

Method

Participants

One hundred and one participants were recruited and tested (aged 20-65 years, mean age = 36 years; 45 female; 94% reporting English as a native language).

Materials and Procedure

Participants were randomly assigned to one of two conditions (rejection, identification) in a be-

tween-subjects design. The procedures and materials were the same as those for Experiment 4, except that the blame statement was slightly adjusted to focus on deserving blame:

> The woman deserves blame for the evaluation she gives the employee.

Results

Independent samples t-tests revealed no effect of condition on any of the three dependent measures. (See Table 12.) Accordingly, the following analyses collapse across condition. Mean response to the could statement was below the midpoint, whereas mean response to the blame statement was above the midpoint. (See Table 13.) A paired samples t-test showed that mean response to the blame statement (M = 4.85, SD = 1.99) exceeded mean response to the could statement (M = 2.61, SD = 2.02), $t(100) = 8.07$, $p < .001$, MD = 2.24, $d = 0.80$.

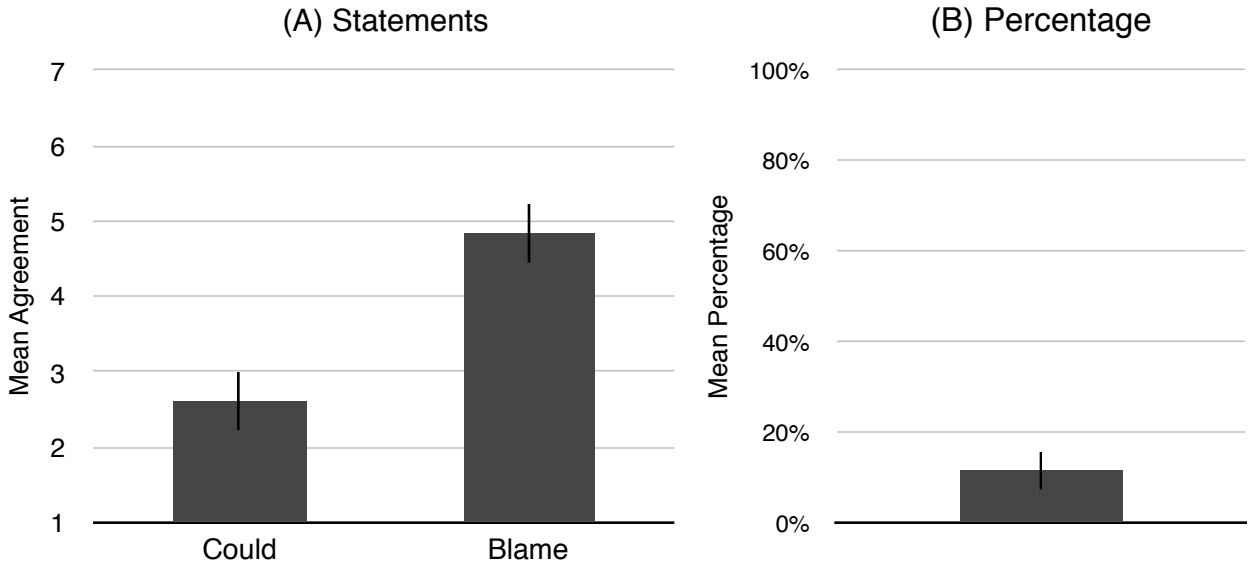

Figure 5. Experiment 5. Panel A: mean response to the test statements about whether the agent could still avoid the outcome, and whether the agent was to blame; the scale ran 1 (SD) - 7 (SA). Panel B: mean estimate of the percentage chance that the agent will avoid the outcome; the scale ran 0-100%. Error bars show 95% confidence intervals. Results collapse across condition (rejection/identification).

Table 12. Experiment 5. Independent samples t-tests comparing responses to the three dependent measures in the rejection and identification conditions.

	Rejection		Identification							
Measure	**M**	**SD**	**M**	**SD**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	2.50	1.93	2.73	2.12	-0.56	99	.577	-0.23	-1.03, 0.58	0.11
blame	4.58	2.08	5.12	1.87	-1.37	99	.175	-0.54	-1.32, 0.24	0.28
percentage	8.90	17.33	13.55	25.66	-1.07	87.92	.288	-4.65	-13.29, 3.99	0.23

Table 13. Experiment 5. One sample t-tests on mean response to the could and blame statements. Test value = 4.

Measure	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	-6.90	100	<.001	-1.39	-1.78, -0.99	0.69
blame	4.31	100	<.001	0.85	0.46, 1.24	0.43

I also analyzed the data from only those participants who answered “strongly disagree” to the could statement and “0%” on the percentage task. In this group (N = 40), mean agreement with the blame statement (M = 4.85, SD = 2.39) was the same as in the sample as a whole and above the midpoint, t(39) = 2.25, p = .030, MD = 0.85, d = 0.36. Modal response was “strongly agree.”

Discussion

People agreed that an agent who could not avoid an outcome still deserved to be blamed for it. This result supports the conclusion that when judging blame in some ordinary cases, people are natural compatibilists. As in Experiment 4, the difference between an agent rejecting and endorsing the outcome upon reflection did not affect blame attributions.

The next experiment investigates whether a similar pattern occurs if participants are asked to judge whether agents *should suffer* for their actions.

Experiment 6

Method

Participants

One hundred and two participants were recruited and tested (aged 18-71 years, mean age = 32 years; 39 female; 96% reporting English as a native language).

Materials and Procedure

Participants were randomly assigned to one of two conditions (rejection, identification) in a between-subjects design. The procedures and materials were the same as those for Experiment 5, except that the blame statement was replaced with this statement:

> The woman should suffer at least some consequences for the evaluation she gives the employee.

Results

Independent samples t-tests revealed no effect of condition on any of the three dependent measures. (See Table 14.) Accordingly, the following analyses collapse across condition. Mean response to the could statement was below the midpoint, whereas mean response to the suffering statement was above the midpoint. (See Table 15.) A paired samples t-test showed that mean response to the suffering statement ($M = 4.68$, $SD = 1.89$) exceeded mean response to the could statement ($M = 2.57$, $SD = 1.83$), $t(101) = 9.40$, $p < .001$, $MD = 2.11$, $d = 0.93$.

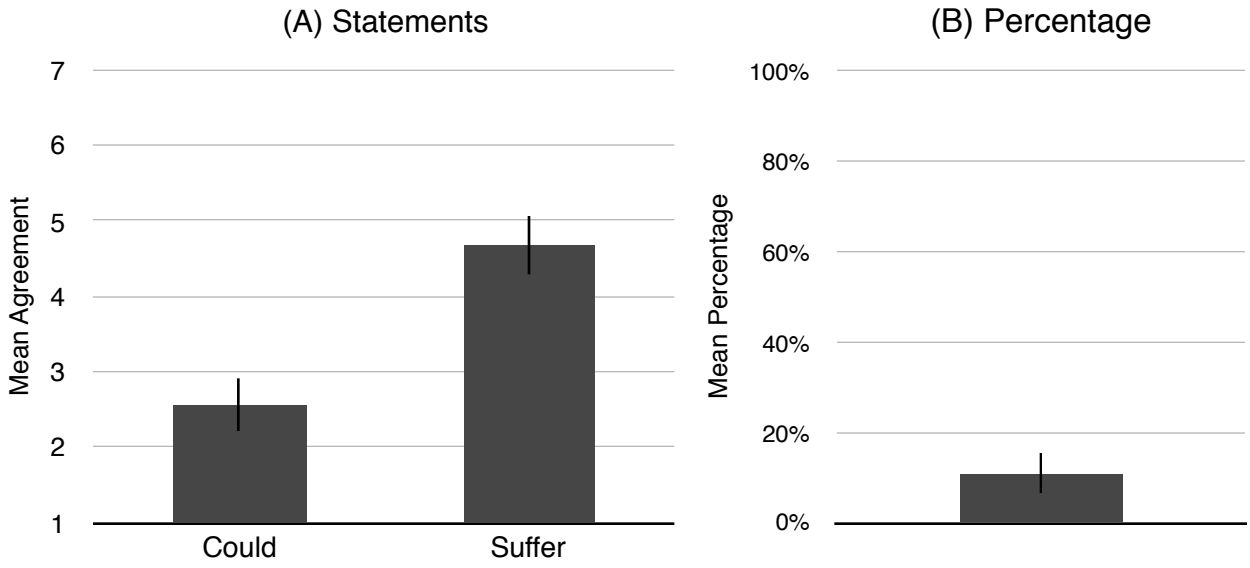

Figure 6. Experiment 6. Panel A: mean response to the test statements about whether the agent could still avoid the outcome, and whether the agent should suffer consequences; the scale ran 1 (SD) - 7 (SA). Panel B: mean estimate of the percentage chance that the agent will avoid the outcome; the scale ran 0-100%. Error bars represent bootstrapped 95% confidence intervals. Results collapse across condition (rejection/identification).

Table 14. Experiment 6. Independent samples t-tests comparing responses to the three dependent measures in the rejection and identification conditions.

	Rejection		Identification							
Measure	**M**	**SD**	**M**	**SD**	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	2.41	1.70	2.73	1.93	-0.87	100	.388	-0.31	-1.03, 0.41	0.17
suffer	4.53	1.82	4.82	1.98	-0.78	100	.436	-0.29	-1.04, 0.45	0.16
percentage	9.98	21.51	11.32	23.22	-0.30	100	.764	-1.34	-10.13, 7.46	0.06

Table 15. Experiment 6. One sample t-tests on mean response to the could and suffering statements. Test value = 4.

Measure	**t**	**df**	**p**	**MD**	**95% CI**	**d**
could	-7.91	101	<.001	-1.43	-1.79, -1.07	0.78
suffer	3.61	101	<.001	0.68	0.30, 1.05	0.36

I also analyzed the data from only those participants who answered "strongly disagree" to the could statement and "0%" on the percentage task. In this group (N = 40), mean agreement with the suffering statement (M = 4.08, SD = 2.29) did not differ from the midpoint, t(39) = 0.21, p = .837, n.s. Modal response was "strongly agree."

Discussion

People agreed that an agent who could not avoid an outcome still should suffer consequences for it. This result supports the conclusion that when judging retribution in some ordinary cases, people are natural compatibilists. As in Experiments 4 and 5, the difference between an agent rejecting and endorsing the outcome upon reflection did not affect attributions.

Conclusion

The results from six experiments support natural compatibilism about a range of categories connected to moral responsibility, including having a moral responsibility to perform an action, being responsible for an outcome, being to blame, deserving blame, and suffering. In response to simple scenarios about familiar activities, people agreed that an agent who could not perform an action still had a moral responsibility to perform it (Experiment 1), was morally responsible for it (Experiment 2), was to blame for it (Experiments 3-4), deserved to be blamed for it (Experiment 5), and should suffer consequences for it (Experiment 6). Except for the last finding on suffering, the same basic patterns persisted among participants who strongly denied, both qualitatively ("strongly disagree") and quantitatively ("0% chance"), that the agent could perform the action or avoid the outcome. Overall, these findings support the conclusion that, over a range of ordinary actions, people can be natural compatibilists about several types of moral responsibility.

The present experiments provide the best evidence to date for natural compatibilism, completely avoiding weaknesses of prior work on the topic (see the Introduction for an explanation of these weaknesses). I used brief, plain, tightly matched, and anodyne stimuli, tested multiple narrative contexts, and included multiple measures to assess how participants understood key variables. Participants understood the stimuli in the relevant way. The manipulations were credible and effective.

The present studies do not show that perceived inability is irrelevant to the moral judgments in question. In fact, quite the opposite is true: people were less likely to attribute the relevant

moral status whenever the agent was unable to perform the relevant action. This suggests that perceived inability lowers our confidence in such attributions, which in turn suggests that incompatibilism is nurtured by some widely shared pre-theoretical intuitions. Nevertheless, in all the studies reported here, the attribution of moral status far exceeded the attribution of the relevant ability. So even if perceived inability moderates moral attributions, it does not provide the hard constraint that natural incompatibilism predicts.

The present results suggest that compatibilism is more natural for some moral categories than others — that is, there might be a spectrum of compatibilist sentiment. In particular, compatibilist judgment seems strongest for *having responsibilities* to do things. This is related to recent findings on the "ought implies can" principle, which show that people attribute moral obligations to agents who cannot fulfill them (Buckwalter & Turri 2014; Buckwalter & Turri 2015; Mizrahi 2015; Chituc, Henne, Sinnott-Armstrong & De Brigard 2016; Turri 2016). The present findings advance understanding of "ought implies can" in commonsense moral psychology by demonstrating that people do not conform to the principle across a wider range of stimuli and procedures. Compatibilist judgment was less strong but still clearly present for *being responsible for* outcomes. Although compatibilist judgment can be comparably strong for *blame*, the present results suggest that people might require more information before attributing blame.

One potential explanation of this need for additional information is that *blame* is a more complicated status than *being responsible for*. There are many ways to elaborate this basic hypothesis. On one version of the hypothesis, *being responsible for* is a two-place relationship between a person and an outcome. It requires only that the agent stand in the relevant relationship

to the outcome, perhaps involving some combination of appropriate causation, intent, and understanding. (The present research was not designed to detect how this relationship is constituted.) By contrast, *blame* is a three-place relation among a person, an outcome, and an audience. When an agent is blameworthy for an outcome, not only is he responsible for it, but the audience should engage in a particular speech act, namely, blaming him for the outcome. The additional logical complexity — namely, being a three-place relation that goes beyond the two-place relation — could explain why participants need more information. For instance, the additional information could be related to the norms of the speech act, whose satisfaction goes beyond whatever is required for agents to be responsible for the outcome.

The present results do not establish that all people are natural compatibilists. There could still be individual differences in this regard and the present research was not designed to rule out this possibility (for evidence that personality traits matter, see Feltz & Cokely 2009). Neither do the results demonstrate that natural compatibilism is the central tendency when evaluating all forms of behavior, all ways of describing behavior as determined, or all categories connected to moral responsibility. For example, I tested scenarios involving interpersonal relations (one person making a promising to or evaluating another), but importantly different patterns might emerge for "self-regarding" actions (Mill 1859), such as the decision to exercise or develop one's talents. To take another example, I did not test scenarios involving overt manipulation, but importantly different patterns might emerge when an agent's inability is due to another person's manipulative activity.

Accordingly, future work could examine in greater detail the extent of natural compatibil-

ism. When probing for the limits of natural compatibilism, researchers should keep in mind the potential for well known biases to mask compatibilist sentiment. In particular, researchers should keep in mind the phenomenon of *excuse validation*, whereby people explicitly misdescribe details of cases involving certain transgressions. More specifically, when evaluating cases of excusable transgressions, roughly half of people deny that a transgression even occurred (Turri 2013; Turri & Blouw 2014). Accordingly, if we treat the inability to fulfill responsibilities as exculpatory, then many of us will likely also be willing to deny responsibility's existence in the first place. Similarly, there might be cases where people think that an agent is responsible for an outcome, but they do not think that he should be blamed for it. In such cases, the desire to excuse could lead people to deny that the agent is even responsible for the outcome. Relatedly, there might be cases where people think that an agent is to blame for an outcome, but they do not think he should suffer anything more severe than verbal criticism or a diminished reputation. In such cases, the desire to excuse the agent from more serious punishment could lead people to deny that the agent is to blame at all. Finally, because suffering consequences is a matter of degree, the desire to excuse the agent from suffering more serious consequences could lead people to deny that the agent should suffering any consequences at all.

It might be wondered whether people attribute the relevant moral status because they believe that at some point in time, not described in the scenario, the agent could have done something that would have prevented his subsequent inability. If so, the objection continues, none of the results would support natural compatibilism. In response, I make three points. First, if natural incompatibilism was true, then it seems unlikely that participants would respond as the objection

envisions. For if the relevant moral status was naturally viewed as incompatible with inability, then it would be misleading to use different time indices for the two judgments, thereby creating pressure to index the two judgments similarly. Second, if participants responded as the objection envisions, the results would still be informative, in two ways. On the one hand, they would still support "narrow" versions of natural compatibilism pertaining to what is possible or unavoidable for the agent in the present context. On the other hand, they would still show that "narrow" compatibilism is more natural for some categories than others. Third, if participants responded as the objection envisions, then it reinforces a possibility noted in the Introduction, namely, that ordinary social cognition might never confront the issue of compatibilism or incompatibilism of any "broader" sort that goes beyond what is explicitly stated and accepted about the current context.

Acknowledgments — For helpful feedback and discussion, I thank Wesley Buckwalter, Ori Friedman, Eddy Nahmias, David Rose, Angelo Turri, and two anonymous referees for this journal. This research was supported by the Social Sciences and Humanities Research Council of Canada, the Ontario Ministry of Economic Development and Innovation, and the Canada Research Chairs program.

References

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness. Personality and

Social Psychology Bulletin, 35(2), 260–268. doi:10.1177/0146167208327217

Buckwalter, W. & Turri, J. (2014). Inability and obligation: compelling counterexamples to 'ought implies can'. Presented at the Buffalo Experimental Philosophy Conference, Buffalo, NY.

Buckwalter, W., & Turri, J. (2015). Inability and Obligation in Moral Judgment. PLOS One, 10(8), e0136589. http://doi.org/10.1371/journal.pone.0136589.g004

Chituc, V., Henne, P., Sinnott-Armstrong, W., & De Brigard, F. (2016). Blame, not ability, impacts moral 'ought' judgments for impossible actions: Toward an empirical refutation of 'ought' implies 'can.' Cognition, 150(C), 20–25. http://doi.org/10.1016/j.cognition.2016.01.013

Cova, F., & Kitano, Y. (2014). Experimental philosophy and the compatibility of free will and determinism: a survey. Annals of the Japan Association for Philosophy of Science, 22, 17–37.

Deery, O., Davis, T., & Carey, J. (2014). The Free-Will Intuitions Scale and the question of natural compatibilism. Philosophical Psychology, 1–26. doi:10.1080/09515089.2014.893868

Feldman, G., Chandrashekar, S. P., & Wong, K. F. E. (2016). The freedom to excel: Belief in free will predicts better academic performance. Personality and Individual Differences, 90(C), 377–383. http://doi.org/10.1016/j.paid.2015.11.043

Feltz, A., & Cokely, E. T. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. Consciousness and Cognition, 18(1), 342–350. doi:10.1016/j.concog.2008.08.001

Frankfurt, H. G. (1988). The importance of what we care about: philosophical essays. Cambridge: Cambridge University Press.

Haidt, J. (2007). The new synthesis in moral psychology. Science, 316(5827), 998–1002. doi: 10.1126/science.1137651

Hume, D. (1748/1993). An enquiry concerning human understanding. (E. Steinberg, Ed.) (2nd ed.). Indianapolis: Hackett.

Kane, R. (1999). Responsibility, luck, and chance: reflections on free will and indeterminism. Journal of Philosophy, 96(5), 217–240.

May, J. (2014). On the very concept of free will. Synthese, 191(12), 2849–2866. doi:10.1007/s11229-014-0426-1

McKenna, M. (2009). Compatibilism. (E. N. Zalta, Ed.) Stanford Encyclopedia of Philosophy. Retrieved December 2014, from http://plato.stanford.edu/entries/compatibilism/

Mill, J. S. (1859/1985). On liberty. London: Penguin Books.

Mizrahi, M. (2015). Ought, can, and presupposition: an experimental study. Methode, 4(6), 232–243.

Monroe, A. E., & Malle, B. F. (2010). From uncaused will to conscious choice: the need to study, not speculate about, people's folk concept of free will. Review of Philosophy and Psychology, 1(2), 211–224. doi:10.1007/s13164-009-0010-7

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: people's psychological concept of free will and its role in moral judgment. Consciousness and Cognition, 27, 100–108. doi:10.1016/j.concog.2014.04.011

Nahmias, E., & Thompson, M. (2014). A naturalistic vision of free will. In E. Machery & E. O'neill (Eds.), Current controversies in experimental philosophy (pp. 86–103). Routledge.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying Freedom: Folk Intuitions about free will and moral responsibility. Philosophical Psychology, 18(5), 561–584. doi:10.1080/09515080500264180

Nahmias, E., Morris, S. G., Nadelhoffer, T., & Turner, J. (2006). Is incompatibilism intuitive? Philosophy and Phenomenological Research, 73(1), 28–53.

Nahmias, E., Shepard, J., & Reuter, S. (2014). It's OK if 'my brain made me do it“: People”s intuitions about free will and neuroscientific prediction. Cognition, 133(2), 502–516. doi: 10.1016/j.cognition.2014.07.009

Nichols, S. (2004). The folk psychology of free will: fits and starts. Mind & Language, 19(5), 473–502.

Nichols, S. (2011). Experimental Philosophy and the Problem of Free Will. Science, 331(6023), 1401–1403. doi:10.1126/science.1192931

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

O'Connor, T. (2010). Free Will. (E. N. Zalta, Ed.) Stanford Encyclopedia of Philosophy. Retrieved December 2014, from http://plato.stanford.edu/entries/freewill/

Pereboom, D. (2001). Living without free will. Cambridge: Cambridge University press.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in Judgments of Moral Blame and

Praise: The Role of Perceived Metadesires. Psychological Science, 14(3), 267–272. http://doi.org/10.1111/1467-9280.03433

Reid, T. (1785). Essays on the active powers of man. Edinburgh: Bell, Robinson & Robinson.

Rose, D., & Nichols, S. (2013). The Lesson of Bypassing. Review of Philosophy and Psychology, 4, 599–619.

Roskies, A., & Nichols, S. (2008). Bringing moral responsibility down to earth. Journal of Philosophy, 105(7), 371–388.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is belief in free will a cultural universal?, 25(3), 346–358.

Schulz, E., Cokely, E. T., & Feltz, A. (2011). Consciousness and Cognition. Consciousness and Cognition, 20(4), 1722–1731. doi:10.1016/j.concog.2011.04.007

Sommers, T. (2010). Experimental Philosophy and Free Will. Philosophy Compass, 5(2), 199–212. doi:10.1111/j.1747-9991.2009.00273.x

Stillman, T. F., Baumeister, R. F., Vohs, K. D., Lambert, N. M., Fincham, F. D., & Brewer, L. E. (2010). Personal philosophy and personnel achievement: belief in free will predicts better job performance. Social Psychological and Personality Science, 1(1), 43–50. doi: 10.1177/1948550609351600

Turri, J. (2013). The test of truth: An experimental investigation of the norm of assertion. Cognition, 129(2), 279–291. doi:10.1016/j.cognition.2013.06.012

Turri, J. (2016). Compatibilism and incompatibilism in social cognition. Cognitive Science. http://doi.org/10.1111/cogs.12372

Turri, J. (in press). Exceptionalist naturalism: human agency and the causal order. Quarterly Journal of Experimental Psychology.

Turri, J., & Blouw, P. (2014). Excuse validation: a study in rule-breaking. Philosophical Studies. doi:10.1007/s11098-014-0322-z

Vihvelin, K. (2011). Arguments for incompatibilism. (E. N. Zalta, Ed.) Stanford Encyclopedia of Philosophy. Retrieved December 2014, from http://plato.stanford.edu/entries/incompatibilism-arguments/

de Waal, F. (2006). Primates and philosophers: how morality evolved. Princeton: Princeton University press.

Watson, G. (1996). Two faces of responsibility. Philosophical Topics, 24(2), 227–248.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. Cognition, 100, 283–301.