Cognition 119 (2011) 166–178

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The paradox of moral focus

Liane Young [a,b,*,1], Jonathan Phillips [c,1]

[a] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, MA, United States
[b] Department of Psychology, Boston College, MA, United States
[c] Program in Cognitive Science and Department of Philosophy, Yale University, CT, United States

ARTICLE INFO

Article history:
Received 18 August 2010
Revised 14 January 2011
Accepted 17 January 2011
Available online 18 February 2011

Keywords:
Morality
Motivated cognition
Force
Focusing
Counterfactual thinking

ABSTRACT

When we evaluate moral agents, we consider many factors, including whether the agent acted freely, or under duress or coercion. In turn, moral evaluations have been shown to influence our (non-moral) evaluations of these same factors. For example, when we judge an agent to have acted immorally, we are subsequently more likely to judge the agent to have acted freely, not under force. Here, we investigate the cognitive signatures of this effect in interpersonal situations, in which one agent ("forcer") forces another agent ("forcee") to act either immorally or morally. The structure of this relationship allowed us to ask questions about both the "forcer" and the "forcee." Paradoxically, participants judged that the "forcer" forced the "forcee" to act immorally (i.e. X forced Y), but that the "forcee" was not forced to act immorally (i.e. Y was not forced by X). This pattern obtained only for human agents who acted intentionally. Directly changing participants' focus from one agent to another (forcer versus forcee) also changed the target of moral evaluation and therefore force attributions. The full pattern of judgments may provide a window into motivated moral reasoning and focusing bias more generally; participants may have been motivated to attribute greater force to the immoral forcer and greater freedom to the immoral forcee.

1. Introduction

When we evaluate moral agents, we consider many factors: whether the agent caused harm (Baron & Ritov, 2004; Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), whether the agent acted with knowledge and intent (Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Cushman, 2008; Malle, 2006; Mikhail, 2007; Young, Nichols, & Saxe, 2010), whether the agent acted freely or under duress or coercion (Darley & Shultz, 1990; Nichols & Knobe, 2007; Woolfolk, Doris, & Darley, 2006), and even what would have happened if the agent had not acted at all (Alicke, Buckingham, Zell, & Davis, 2008; Branscombe, Owen, Garstka, & Coleman, 1996; Nario-Redmond & Branscombe, 1996). Ordinary folk and legal scholars alike take these to be relevant factors for moral judgment (Hart, 1968; Mikhail, 2007).

A growing body of evidence, however, shows that moral judgments can influence our evaluations of these very factors (Knobe, 2010; Pettit & Knobe, 2009). When we judge an agent to have acted immorally, we are subsequently more likely to judge the agent to have caused the bad outcome (Alicke, 2000; Cushman, Knobe, & Sinnott-Armstrong, 2008; Hitchcock & Knobe, 2009; Knobe & Fraser, 2008), to have acted with knowledge and intent (Beebe & Buckwalter, 2010; Knobe, 2003), to have acted freely rather than under force (Harvey, Harris, & Barnes, 1975; Phillips & Knobe, 2009), and we are even more likely to consider counterfactual events, that is, what would have happened if the agent had not acted immorally (McCloy & Byrne, 2000; N'gbala & Branscombe, 1997; Roese, 1997).

While this research has demonstrated the widespread impact of moral judgment on non-moral features of

* Corresponding author at: Department of Brain and Cognitive Sciences, 46-4021, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, United States.
E-mail address: liane.young@bc.edu (L. Young).
[1] These authors contributed equally to this manuscript.

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.01.004

cognition, important questions remain unanswered: Why and how does this influence occur? Some researchers have claimed that morality is actually a core component of the concepts of causation, knowledge, intent, freedom, and counterfactual thinking (Pettit & Knobe, 2009; Phelan & Sarkissian, 2008; Wright & Bengson, 2009). Others suggest that this influence reflects motivated moral reasoning, whereby non-moral attributions are altered either consciously or unconsciously to support initial moral judgments (Adams & Steadman, 2004; Alicke, 2008; Nadelhoffer, in press). The present research proposes to inform this debate by investigating an unexplored implication of the prior literature.

Previous research has established that how we evaluate an agent in moral terms affects our non-moral evaluations of that agent. Yet, when bad outcomes occur in ordinary, everyday life, more than one moral agent is often responsible (Gray & Wegner, 2009; Waytz, Gray, Epley, & Wegner, 2010). For example, when prisoners of war are tortured, a commander may have issued the order, which was then carried out by subordinate soldiers. The combination of the fact that moral judgments influence non-moral cognition and the fact that multiple agents are often morally responsible for a single outcome reveals an important implication of the prior literature: shifting the blame from one agent to another agent within a single scenario could lead observers to make contradicting non-moral attributions. For example, if observers focus on the *soldiers* as the morally blameworthy agents, they might judge that the soldiers freely tortured the prisoners, that they were not truly forced by the commander, and that they could have resisted the order. However, if observers redirect their focus to the *commander*, then they might judge that the commander exerted great force on the soldiers to get them to behave badly and that he gave them no other option.

We propose that this paradox is the consequence of *moral focus*: when observers judge an agent to have acted immorally, they focus on that immoral agent, which consequently alters their non-moral evaluations of that agent. By investigating this phenomenon, we aim to illuminate the role of moral blame, the underlying mechanism of focusing, and, importantly, the nature and boundaries of the broader influence of moral judgments on non-moral attributions.

1.1. Focusing bias and counterfactual thinking

Focusing in reasoning and decision-making more generally has been established in extensive research on mental models (Legrenzi, Girotto, & Johnson-Laird, 1993). In a series of studies, individuals were shown to focus only on factors explicitly represented in their mental models, leading to errors in logic (e.g., modus tollens, Wason's selection task), riskless decisions, and counterfactual reasoning. In addition, these studies revealed that participants focus particularly on information about the protagonist, and events from the protagonist's perspective, including the protagonist's actions (i.e. what the protagonist did), and counterfactual actions (i.e. what the protagonist could have done but did not do). Focusing on this information, explicitly represented in mental models, resulted in departures from rational principles, i.e. "focusing bias".

Related research has explored the relationship between counterfactual reasoning and moral judgment. In a pair of studies by Branscombe and colleagues, directing participants to engage in counterfactual thinking, that is, imagining alternatives to reality (e.g., "he might have behaved better"/"she could have acted differently") affected moral judgments of victims and assailants in the case of rape (Branscombe et al., 1996; Nario-Redmond & Branscombe, 1996). In one case, participants considered counterfactuals in which the outcome was worse than the original event. When they focused on how the *assailant's* behaving differently could have brought about a worse outcome, moral condemnation of the crime decreased. However, when they focused on how the *victim's* behaving differently could have brought about a worse outcome, condemnation of the crime increased. Moral judgment is therefore influenced by participants' focus on one agent or another in their counterfactual thinking. A related series of studies targeting the opposite influence of moral judgment on counterfactual thinking found that participants engaged in more counterfactual thinking for "immoral" as opposed to neutral causes of an outcome (N'gbala & Branscombe, 1997). Together, these studies suggest the complex relationship between focusing, counterfactual thinking, and moral judgment.

The current study tests the specific hypothesis that moral focus – focus on one agent or another as the principal moral agent – leads to logical inconsistencies in subsequent non-moral judgments about the agents. Moral evaluation of an agent should motivate participants to focus on that agent and modify other judgments of that agent, like whether the agent could have done otherwise. Consequently, this motivated reasoning should be consistent with participants' moral judgments but, as a result, reflect internal errors, as observed in focusing bias more generally.

1.2. Force and freedom

To investigate the role of moral focus, we use judgments of *force* as a case study. Consider the important relationship between moral judgment and judgments of force. When moral agents cause harm, evaluations of those agents depend critically on how the harm was caused: freely or under duress or coercion. Important work in social psychology over the last five decades has helped us understand just how free or forced the average agent may be in certain situations, and also how observers perceive those agents. Ordinary experimental participants have been shown to knowingly harm others under the right kinds of force, including authority and consensus (e.g., Asch, 1951; Milgram, 1963; Ross, 1988; Zimbardo, 1973). These findings came as an initial surprise because people tend to over-attribute "freedom" to agents more generally; for example, participants spontaneously attribute internal traits to individuals who are instructed by the experimenter to write counter-attitudinal essays (Jones & Harris, 1967; Snyder & Jones, 1974). Observers' tendency to underestimate situational forces (e.g., authority) and to assume that people have the ability to resist force often leads to greater moral condemnation of the "forcee" when he or she is forced to behave badly. In sum, judgments of whether an agent was truly forced to

do harm, or whether the agent acted freely, crucially influence people's moral judgments of that agent.

Is it also the case that moral judgments influence people's judgments of force and freedom? Specifically, when an agent is forced to perform an *immoral* action, are observers even more likely to judge that the agent did so freely? Recent research has shown that moral evaluations influence attributions of force in exactly this way (Phillips & Knobe, 2009). Specifically, agents acting under a fixed level of situational constraint are evaluated as less forced to perform immoral actions than actions which are not immoral. For example, participants read one story set in a hospital, in which the chief of surgery orders a doctor, against the doctor's will, to prescribe a drug that will save a patient's life. Participants were then asked the key question: "Was the doctor forced to prescribe the drug?" On average, participants judged that the doctor was indeed forced to prescribe the drug. However, when participants instead read a story in which the chief orders the doctor, against the doctor's will, to prescribe a drug that will *kill* the patient, participants gave the opposite answer, judging that the doctor was not forced to prescribe the drug; he was free to do otherwise – that is, to do the right thing.

Suppose we re-frame the question, but redirect participants' focus from the doctor to the chief, and ask: "Did the chief force the doctor to prescribe the drug?" This question bears a logical relation to the first. The statement "the chief forced the doctor" entails that "the doctor was forced by the chief". However, in spite of this relation, these questions lead one to focus on different moral agents, making one agent more salient than the other.

The interpersonal structure of the doctor–chief scenario above provides a unique opportunity to investigate the impact of moral judgment on non-moral attributions. One possibility is that, since people judged that the doctor was not forced to act immorally, they will also judge that the chief did not force the doctor to act immorally, on logical grounds (i.e., Y was not forced by X, and X did not force Y). But, alternatively, people could judge that, though the doctor was not forced by the chief to act immorally, the chief did in fact force the doctor to act immorally – a paradoxical pattern (i.e., X forced Y, but Y was not forced by X). On this alternative, which we favor, participants may be motivated, consciously or unconsciously, to attribute more or less force depending on the moral agent in focus – resulting in logical inconsistencies[2] in the non-moral domain of force. Put plainly, participants' moral focus will alter their force judgments in a way that supports their moral judgments.

We conducted five experiments to investigate the impact of moral evaluations on force attributions. Experiments 1 and 3 replicated the previously observed asymmetry in forcee judgments (Phillips & Knobe, 2009) and also revealed the predicted paradoxical pattern by comparing forcee to forcer judgments: participants judged that the forcee was not forced to act immorally, but that the forcer forced the forcee to act immorally. These effects were found to be specific to human agents versus physical forces (Experiment 2), and specific to cases in which bad outcomes were brought about knowingly versus unknowingly (Experiment 4). Finally, we demonstrated that directly manipulating participants' focus of attention on one moral agent versus another (forcer versus forcee) changed the putative target of moral evaluation and therefore participants' force attributions (Experiment 5).

2. Experiment 1: the paradox

Experiment 1 presents a replication of the asymmetry in forcee judgments observed in previous research (i.e. Y was forced by X to act morally but not immorally) but in addition elicits judgments about the forcer for the same scenarios: Did X force Y to act? Experiment 1 uses an interpersonal relationship (i.e., the captain of a ship and a sailor under his command), and compares morally bad behavior (throwing passengers overboard) to morally neutral behavior (throwing cargo overboard). We hypothesized that when participants were asked about the sailor ("forcee"), they would judge that he was *not* forced to act immorally, as in prior work, but when participants were asked about the captain ("forcer"), they would judge that he *did* force the sailor to act immorally. That is, participants' judgments of force would be determined by their focus on either the captain or the sailor as the relevant moral agent.

2.1. Method

We collected data from 120 participants, using the web-based resource Amazon Mechanical Turk (http://www.mturk.com/). Three measures were taken to screen out repeat participants on a 7-point scale enchored at "not at all" (1) and "absolutely" (7): (1) we asked that people not participate if they had previously taken a similar survey, (2) participants answered a final question about whether they had completed a similar survey before and, if so, its topic, (3) we eliminated data from participants with identical worker IDs. We eliminated 11 participants. Participants were paid $0.10 for approximately one minute of their time.

Participants read one of two versions (morally neutral or morally bad)[3] of the following scenario (Phillips & Knobe, 2009):

> While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the

[2] In making the claim that this pattern of responses is formally inconsistent, we are relying on the premise that "X forces Y to do p" entails "Y is forced to do p". We also note that while statements about 'force' may have a *performative* reading (i.e., "X *attempted* to force Y to do p") and a *success* reading (i.e., "X *succeeded* in forcing Y to do p"), all statements about 'force' in the current study concerned scenarios in which Y successfully forced X to do p. Thus, we find alternative explanations which rely on this distinction unlikely in that they all have to assume that participants systematically relied on a performative reading for statements of the form 'X forced Y to do p' and systematically relied on a success reading for statements of the form 'Y was forced to do p' but only did so in scenarios involving immoral actions and despite the fact that the forcing was successful.

[3] To validate this condition manipulation, we collected moral judgment data from a separate group of 83 participants. As in previous research (Phillips & Knobe, 2009), participants judged throwing passengers overboard to be morally worse than throwing cargo overboard when asked about both the captain ($t(81) = 22.3$, $p < 0.001$) and the sailor ($t(81) = 16.4$, $p < 0.001$).

captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to throw his *expensive cargo*/*passengers* overboard. Thinking quickly, the captain ordered one of his sailors to throw the *cargo*/*passengers* overboard. While the *cargo*/*passengers* sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

Participants were then asked one of the following two questions on a 7–point scale anchored at "not at all" (1) and "absolutely" (7): (1) Sailor: Was the sailor forced to throw the *cargo*/*passengers* overboard? (2) Captain: Did the captain force the sailor to throw the *cargo*/*passengers* overboard? Accordingly, participants were assigned to one of four conditions in a 2 × 2 experimental design: (1) participants responded to either the morally neutral or morally bad scenario (cargo versus passengers), and (2) participants made a judgment focused on either the forcer (whether the *captain* forced the sailor), or the forcee (whether the *sailor* was forced by the captain). Critically, we predicted an interaction between moral valence (neutral versus bad) and focus (sailor versus captain). First, the captain (forcer) should be judged as more forceful for bad versus neutral outcomes, while the sailor (forcee) should be judged as less forced for bad versus neutral outcomes. Second, for bad outcomes only, and not neutral outcomes, participants should judge the captain (forcer) as forcing the sailor (forcee) more than the sailor was forced by the captain.

In addition to these primary analyses, we collected additional data from a new group of 83 participants to investigate whether participants might be interpreting a term like "force" differently for the forcer versus the forcee, since such a difference could account for any observed discrepancy between force judgments of the forcer and forcee. Previous work has shown a distinction between an agent who 'causes' an outcome and one who 'enables' an outcome (Frosch, Johnson-Laird, & Cowley, 2007). Consider the following illustration in which Mary *enables* and Laura *causes* an outcome (Frosch et al., 2007): "Mary threw a lighted cigarette into a bush. Just as the cigarette was going out, Laura deliberately threw petrol on it. The resulting fire burnt down her neighbor's house." Previous work suggests that Laura, who caused the outcome, is judged as more responsible for what happened and more deserving of punishment (e.g., prison time, damages), than Mary, who only enabled the outcome. This distinction could also apply to the above scenario in that participants might view the captain or the sailor as "causing" versus merely "enabling" an outcome. For the present experiment, we provided participants with the example above of causing versus enabling, and then we asked participants whether they saw each agent (captain, sailor) as a cause or an enabler; half of the participants read the immoral scenario (throwing passengers), and the other half the neutral scenario (throwing cargo).

2.2. Results and discussion

As predicted, a 2 (moral valence: morally neutral versus morally bad) × 2 (focus: sailor versus captain) between-subjects ANOVA revealed the key interaction between moral valence and focus ($F(1, 108) = 7.14$, $p = 0.009$, partial $\eta^2 = 0.06$), and no main effects (Fig. 1).

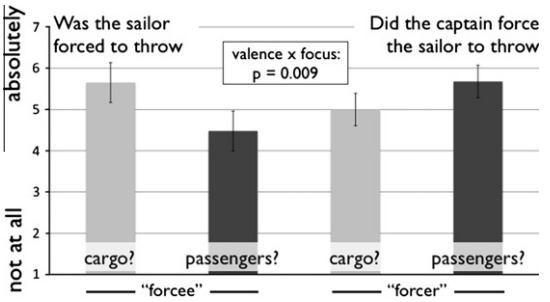

Fig. 1. Force judgments of the forcee (sailor, left) and the forcer (captain, right) for neutral actions (light bars) versus morally bad actions (dark bars). Error bars indicate standard error.

First, as predicted, when asked about the sailor, participants judged that the sailor was forced less to throw the passengers overboard (immoral) than the cargo (neutral) ($t(49) = 2.15$, $p = 0.037$), replicating prior work (Phillips & Knobe, 2009). However, as reflected in the significant interaction between moral valence and focus, we found the opposite trend when participants were asked instead about the captain: participants judged that the captain forced the sailor more to throw the passengers overboard (immoral) than the cargo (neutral), though this trend was not significant ($t(56) = 1.55$, $p = 0.128$). Second, as predicted, participants judged that the captain forced the sailor to throw the passengers overboard more than the sailor was forced to throw the passengers ($t(44) = 2.22$, $p = 0.032$). This discrepancy emerged only for the morally bad scenarios, not for the neutral scenarios. This paradoxical pattern, reflected in the key interaction, therefore reveals a discrepancy in participants' force judgments of immoral actions – whether the captain forced the sailor to throw the passengers (i.e. X forced Y) and whether the sailor was forced by the captain to throw the passengers (i.e. Y was not forced by X).

Finally, participants' judgments of whether an agent served as a cause or an enabler revealed that, on average, participants saw both the captain (51 out of 83 participants; $\chi^2(1, N = 83) = 4.3$, $p = 0.037$) and the sailor (53 out of 83 participants; $\chi^2(1, N = 83) = 6.4$, $p = 0.01$) as causes rather than enablers. Importantly, participants did not distinguish between the captain and the sailor in the extent to which they saw them as causes versus enablers overall ($Z = -0.295$, $p = 0.77$), or separately for immoral scenarios ($Z = -0.218$, $p = 0.83$) and moral scenarios ($Z = -0.600$, $p = 0.55$). There was also a trend such that, overall, agents were judged as causes versus enablers more for immoral versus moral scenarios ($\chi^2(2, N = 83) = 5.13$, $p = 0.077$); that is, subjects were more likely to see an *immoral* agent as a cause versus an enabler, compared to a neutral agent. This is consistent with prior work showing that agents who are causes of bad outcomes are judged as more responsible and more deserving of punishment than enablers (Frosch et al., 2007). In sum, these analyses suggest that both the captain and the sailor are similarly seen as causes of bad outcomes (e.g., passengers thrown overboard).

3. Experiment 2: agents versus non-agents

The ship scenarios of Experiment 1 enable further investigation of the observed asymmetry in force attributions for bad versus good outcomes. Here we hypothesized that this asymmetry will arise only for cases in which an *agent* does something immoral. Specifically, participants should not attribute more force to a storm (non-agent), when the storm is described as causing passengers versus cargo to be thrown overboard. Instead, participants should focus on the immoral actions of the sailor who carried out the action, and, as a result, attribute even less force to the storm in the case of a bad outcome. This hypothesis is based on the premise that, in general, participants should be willing to assign moral blame to agents, but not non-agents, for causing bad outcomes. Accordingly, participants should not attribute greater force to non-agents for causing bad versus neutral outcomes. The key analysis of this experiment therefore concerns the comparison between the force of the captain (agent) on the sailor versus the force of the storm (non-agent) on the sailor, in the case of bad and neutral outcomes.

3.1. Method

We collected data from 120 new participants, and eliminated four repeat participants. Participants read the same two versions (bad versus neutral outcomes) of the ship scenario from Experiment 1. Participants made attributions of force to the storm: Did the *storm* force the sailor to throw the cargo/passengers overboard?[4] We compared these new judgments to the original judgments collected in Experiment 1 in response to the captain question: Did the *captain* force the sailor to throw the cargo/passengers overboard? Critically, to compare force attributions to an agent versus a non-agent, we used a 2 (outcome valence: bad versus neutral) × 2 (agency: did the captain force the sailor? versus did the storm force the sailor?) design. We predicted an interaction between outcome valence (neutral versus bad) and agency (captain versus storm). First, the captain (agent) should be judged as more forceful than the storm (non-agent) only in the case of bad outcomes, not neutral outcomes. Second, the captain (agent) but not the storm (non-agent) should be judged as more forceful for bad versus neutral outcomes.

3.2. Results and discussion

To directly compare force attributions to the agent (i.e. did the captain force the sailor?) versus non-agent (i.e. did the storm force the sailor?), we conducted a 2 (outcome valence: bad versus neutral) × 2 (agency: captain versus storm) between-subjects ANOVA. We observed both main effects of outcome valence ($F(1, 115) = 5.04$, $p = 0.027$, partial $\eta^2 = 0.04$) and agency ($F(1, 115) = 14.9$, $p < 0.001$, partial $\eta^2 = 0.12$) and the predicted interaction between outcome valence and agency ($F(1, 115) = 63.2$, $p < 0.001$, partial $\eta^2 = 0.13$) (Fig. 2.). Participants attributed more force to the captain (agent) versus the storm (non-agent) overall, as reflected in the main effect of agency. Participants also attributed more force in the case of neutral outcomes (throwing cargo) versus bad outcomes (throwing passengers), as reflected in the main effect of outcome. Critically, though, both of these main effects were observed in the context of the key interaction between agency and outcome, which we unpack below.

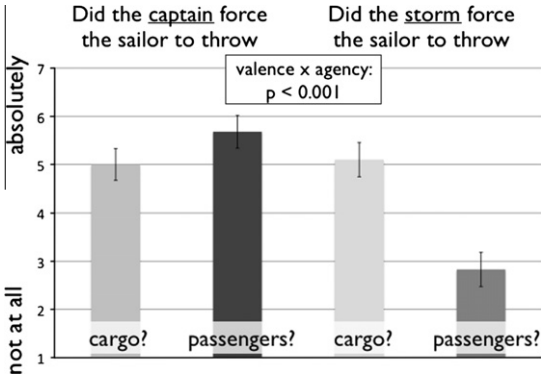

Fig. 2. Force judgments of the agent (captain, left) and the non-agent (storm, right) for neutral actions (light bars) versus morally bad actions (dark bars). Error bars indicate standard error.

First, as predicted, it was only in the case of a bad outcome (not a neutral outcome) that participants judged the captain as applying more force than the storm; in particular, participants judged that the *captain* forced the sailor to throw the passengers overboard more than they judged that the *storm* forced the sailor to throw the passengers overboard ($t(55) = 5.79$, $p < 0.001$). However, in the neutral case (e.g., throwing cargo), participants did not distinguish between the storm and the captain. Therefore, participants selectively attributed more force to *agents* who acted *immorally*.

Second, as predicted, and as reflected in the agency by outcome interaction, while participants attributed less force to the storm when passengers (versus cargo) were thrown overboard, participants showed the opposite trend for the captain, as reported in Experiment 1: participants attributed more force to the captain when passengers (versus cargo) were thrown overboard. Moreover, participants appeared especially unwilling to attribute force to the storm in the case of a bad outcome, perhaps even more so because an agent (the sailor) was available for moral blame. Participants may be very willing to blame agents for a bad outcome but less willing to blame non-agents for the same outcome.

These results show that participants attribute greater force to *agents* for bringing about bad outcomes;

[4] For completeness, we actually collected responses to two questions: (1) Did the storm force the *captain* to throw the cargo/passengers overboard? (2) Did the storm force the *sailor* to throw the cargo/passengers overboard? We predicted and found no difference between these questions. A 2 (outcome valence: bad versus neutral) × 2 (question: storm's force on captain versus storm's force on sailor) between-subjects ANOVA yielded a main effect of outcome valence ($F(1, 115) = 47.3$, $p < 0.001$, partial $\eta^2 = 0.30$), no main effect of question, and no interaction between outcome valence and question, indicating no difference between the two questions. As reflected in the main effect of outcome valence, participants judged that the storm forced both the sailor and the captain more to throw the cargo (mean force judgment: 5.39, standard error: 0.26) versus the passengers (mean: 2.83, standard error: 0.27).

participants are willing to blame agents, but not non-agents, for causing bad outcomes. Participants may even reduce the amount the force they attribute to non-agents when agents are available for moral focus.

4. Experiment 3: moral obligation

Experiment 3 extends the results of the previous experiments in two important ways. First, Experiment 3 uses another instance of an interpersonal relationship – a chief of surgery who forces a doctor either to kill a patient the doctor likes or to save a patient the doctor dislikes.[5] Second, Experiment 3 investigates the comparison between morally bad behavior (e.g., killing the patient) and morally obligatory behavior (e.g., saving the patient), rather than the comparison between morally bad behavior (e.g., throwing passengers overboard) and morally neutral behavior (e.g., throwing cargo overboard). Two hypotheses for Experiment 3 follow.

First, we hypothesized that participants' judgments would reflect the same paradox observed in Experiment 1 for immoral actions: when asked about the forcee (i.e. the doctor), participants should judge that the doctor was not forced to act immorally, that is, to kill the patient, but when asked about the forcer (i.e. the chief), participants should judge that the chief did force the doctor to act immorally. Participants' force judgments should therefore be determined by their focus on either the chief or the doctor as the primary moral agent. Participants should then deliver force judgments that are in line with their assignments of moral blame to the agent in focus.

Second, we hypothesized that the opposite pattern would obtain for the morally good or obligatory action of saving the patient's life: participants should judge that the chief did not force the doctor to act morally, that is, to save the patient's life (X did not force Y), but that the doctor was indeed forced by the chief to save the patient's life (Y was forced by X). Notably, the scenario states that the doctor does not want to save the patient because he dislikes her. Given this negative description, participants should be motivated to make negative moral judgments of the doctor and consequently deliver force judgments that support these negative moral judgments; specifically, participants should judge that the doctor did not save the patient freely but was forced by the chief to do so. It is also possible that forcers are not seen as having to apply as much force to get forcees to do the right thing, what is already morally (or legally) obligatory. Thus, the chief may be judged as less forceful when he forces the doctor to save (versus kill) the patient.

[5] The same pattern of results was replicated using a scenario based on war crimes that have actually occurred. After reading about a military unit commander who ordered his soldiers to brutally torture captured enemy rebels as a method of obtaining information, participants were asked to rate their agreement with one of two possible sentences: (1) 'The military commander forced his soldiers to brutally torture the rebels' or (2) 'The military soldiers were forced to brutally torture the rebels by the unit commander'. Participants judged that forcer (the unit commander) forced the forcee (the soldiers) (mean = 5.54) more than the soldiers were forced by the unit commander (mean = 4.53; $t(75) = 2.60$, $p = .011$).

4.1. Method

We collected data from 100 new participants, and eliminated four repeat participants. Participants were assigned randomly to one of four conditions, in the same 2×2 experimental design as Experiment 1. Participants read one of two versions of the following scenario (Phillips & Knobe, 2009):

> At a certain hospital, there were very specific rules about the procedures doctors had to follow. The rules said that doctors didn't necessarily have to take the advice of consulting physicians but that they did have to follow the orders of the chief of surgery. One day, the chief of surgery ordered a doctor to prescribe the drug Accuphine for a patient. The doctor *had always disliked this patient and actually didn't want her to be cured/really liked the patient and wanted her to recover as quickly as possible*. However, both the doctor and the chief knew that giving this patient Accuphine would result in *an immediate recovery/her death*. The doctor went ahead and prescribed Accuphine. The patient *recovered immediately/died shortly thereafter*."

Participants were asked one of two force questions on a 7-point scale anchored at "not at all" (1) and "absolutely" (7): (1) Doctor: Was the doctor forced to prescribe Accuphine? (2) Chief: Did the chief of surgery force the doctor to prescribe Accuphine?

As in Experiment 1, we predicted an interaction between moral valence (good versus bad) and focus (doctor versus chief). First, the chief (forcer) should be judged as more forceful for bad versus good outcomes, while the doctor (forcee) should be judged as less forced for bad versus good outcomes. Second, for bad outcomes, participants should judge the chief (forcer) as forcing the doctor (forcee) more than the doctor was forced by the chief, but for good outcomes, participants should judge the chief as forcing the doctor less than the doctor was forced by the chief.

4.2. Results and discussion

Participants judged whether the doctor was forced by the chief to save the patient (morally good/obligatory) versus kill the patient (morally bad), and whether the chief forced the doctor to save versus kill the patient. A 2 (outcome valence: good versus bad) $\times$ 2 (focus: doctor versus chief) between-subjects ANOVA yielded the predicted interaction between outcome valence and focus ($F(1, 99) = 13.3$, $p < 0.001$, partial $\eta^2 = 0.12$), and no main effects (Fig. 3).

First, as in previous research (Phillips & Knobe, 2009), we found a trend in participants' judgments that the doctor was forced less to kill than to save ($t(58) = 1.93$, $p = 0.058$). However, when asked about the chief instead of the doctor, participants judged that the chief forced the doctor more to kill than to save ($t(38) = 3.48$, $p = 0.001$). In other words, though both scenarios described the doctor as being forced to act against his will, participants judged that the doctor was forced less to kill the patient than to save the patient. By contrast, though both scenarios described the chief as

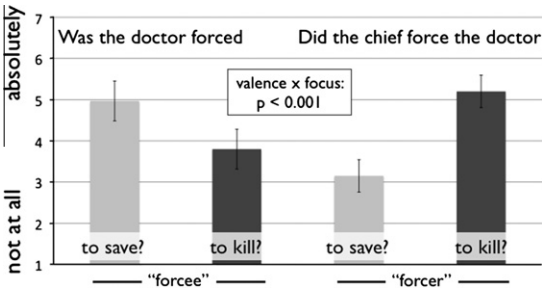

Fig. 3. Force judgments of the forcee (doctor, left) and the forcer (chief, right) for good outcomes (light bars) versus bad outcomes (dark bars). Error bars indicate standard error.

forcing the doctor to act against his will, participants judged that the chief forced the doctor *more* to kill the patient than to save the patient.

Second, participants also judged that the chief forced the doctor to kill more than the doctor was forced by the chief to kill ($t(48) = 2.43$, $p = 0.019$). As predicted, the opposite inconsistency emerged for the morally good (obligatory) scenario: participants judged that the doctor was forced by the chief to save, but that the chief did not force the doctor to save ($t(48) = 2.90$, $p = 0.006$). The chief may have been seen as not having to *force* the doctor to fulfill a moral or legal obligation. More importantly, the scenario presents the doctor as having "always disliked this patient" and not wanting her cured. Participants may have been reluctant to judge the doctor as having freely done the right thing in order to support their negative moral judgment of the doctor. Therefore, participants may have been motivated to judge that the doctor was *forced* to save the patient. Whether this pattern would obtain even if the doctor had not initially been described in a negative light is worth exploring. Notably, though, describing the forcee as needing to be *forced* to do the right thing may necessarily put the forcee in a negative moral light, thereby motivating participants to judge him harshly and alter their force judgments.

In sum, the results suggest that participants engaged in motivated moral reasoning for both morally bad and morally obligatory scenarios, resulting in inconsistent force attributions.

5. Experiment 4: moral valence versus outcome valence

In Experiment 3, participants judged that the chief forced the doctor less to save the patient than to kill the patient. Experiment 4 tests whether this asymmetry depends on the chief's moral status or simply on the valence of the outcome brought about by a moral agent. Ignorance, or a mistake of fact, is often treated as a mitigating factor in moral judgments (e.g., murder versus manslaughter) (Cushman, 2008; Hart, 1968; Mikhail, 2007; Young & Saxe, 2009). In Experiment 3, the chief acted with full knowledge of the drug's effect. Therefore, to explore the nature of the observed asymmetry in attributions of force for killing versus saving, we modified the scenarios such that the chief was ignorant of the drug's effects on the patient.

5.1. Method

We collected data from 40 new participants; we eliminated three repeat participants. Participants read one of two versions (e.g., good versus bad outcome) of a modified scenario, replacing the following sentence, "However, both the doctor and the chief knew that giving this patient Accuphine would result in an immediate recovery/her death", with the sentence, "*However, the doctor – but not the chief – knew* that giving this patient Accuphine would result in an immediate recovery/her death". All participants were asked: Did the chief of surgery force the doctor to prescribe Accuphine? We predicted an interaction between valence (bad versus good) and knowledge (knowing versus ignorant).

5.2. Results and discussion

To analyze judgments for the original conditions (i.e. knowing chief) and the new conditions (i.e. ignorant chief), we conducted a 2 (outcome valence: bad versus good) $\times$ 2 (knowledge: knowing versus ignorant) between-subjects ANOVA, yielding a significant interaction between outcome valence and knowledge ($F(1, 76) = 10.0$, $p = 0.002$, partial $\eta^2 = 0.12$), and no main effects (Fig. 4).

Independent-samples t-tests enabled a separate analysis of the new conditions. When participants were asked whether the *ignorant* chief forced the doctor to prescribe the drug, participants did not distinguish between the bad outcome (mean: 3.26, standard error: 0.50) and good outcome (mean: 4.17, standard error: 0.54; $t(35) = 1.23$, $p = 0.23$). The absence of a difference between attributions of force for the "*unknowing* killer" and the "*unknowing* saver" contrasts with the difference observed for the chief who acted knowingly.

In addition, we observed a difference between force attributions to the knowing versus ignorant chief only for the killing case ($t(33) = 3.12$, $p = 0.004$), but not for the saving case ($t(36) = 1.44$, $p = 0.16$) (Fig. 4), broadly in line with the effects observed in Experiment 2: participants distinguished between the agent and non-agent in the case of a bad outcome only (throwing passengers). Together, these results suggest that participants may treat ignorant agents

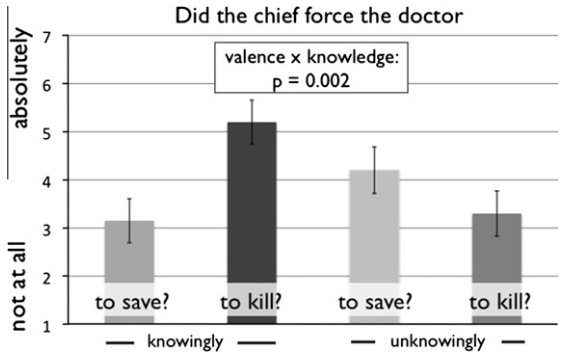

Fig. 4. Force judgments of the chief when he knowingly (left) versus unknowingly (right) forces the doctor to bring about a bad outcome. Error bars indicate standard error.

(i.e. ignorant chief) and non-agents (i.e. storm) similarly, as compared to knowing agents.

In sum, participants did not attribute more force to knowing versus unknowing agents in general, but specifically for agents who knowingly caused harm and were therefore deserving of moral blame. Experiment 4 suggests that the asymmetry in force attribution observed in the previous experiments is specific to moral evaluations of agents rather than the mere valence of outcomes brought about by the agents. When one agent forces another to do something he *knows* to be harmful, he is judged as more forceful.

6. Experiment 5: moral focus

Why would asking about the forcer (e.g., chief, captain) versus the forcee (e.g., doctor, sailor) produce inconsistent patterns of force judgments in Experiments 1–4? We hypothesized that asking about the forcer focuses participants' attention on the moral status of the forcer, whereas asking about the forcee focuses attention on the moral status of forcee. When participants are focused mainly on the forcer (e.g., captain) as the moral agent, then they may be motivated to make force judgments that support the conclusion that the forcer is to blame for the bad outcome – attributing more force to the captain. But when participants are focused mainly on the forcee (e.g., sailor) as the moral agent, they may be motivated to attribute more freedom to the forcee to support the conclusion that he is to blame for the bad outcome.

If participants are motivated consciously or unconsciously to change their force judgments to support their moral judgments, then changing the focus of those moral judgments should also change their force judgments. Experiments 1–4 accomplished this by asking participants to make an explicit judgment about either the forcer or the forcee. In Experiment 5, we took a different approach – directing participants' focus to either the forcer or the forcee in the text of the scenario itself. In particular, we hypothesized that increased focus on the sailor and decreased focus on the captain would reduce force attributions to the captain.

6.1. Method

We collected data from 56 new participants. Participants made force judgments for the captain for one of two scenarios: (1) the version of the ship scenario that focused on the captain, as used previously, in brackets, or (2) a new version that focused on the sailor:

> While he was sailing on the sea, a large storm came upon a *sailor [captain]* on a ship. The waves began to grow larger, and the *sailor's [captain's]* small vessel was too heavy. The *sailor's [captain's]* ship would flood and the *sailor [captain]* would drown if he didn't make it lighter. The only way that the *sailor [captain]* could keep the ship from capsizing was to throw the passengers overboard. Thinking quickly, the captain of the ship ordered the sailor to throw the passengers overboard. The sailor threw the passengers overboard. While the passengers sank to the bottom of the sea, the *sailor [captain]* was able to survive the storm and returned home safely.

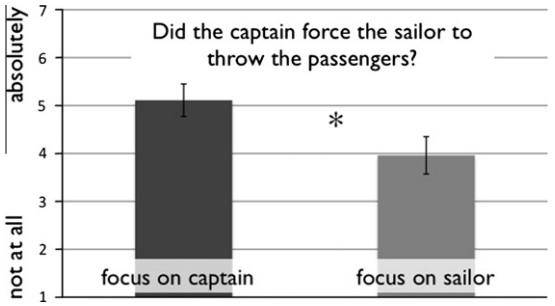

Fig. 5. Force judgments of the captain for scenarios that focus attention on the captain (left) versus the sailor (right). Error bars indicate standard error.

All participants were asked the same question: Did the captain force the sailor to throw the passengers overboard? To ensure that participants read the scenario carefully, they were also asked a control question: Did the captain tell the sailor to throw the passengers overboard? Three participants were excluded from the analysis because they failed the control question, answering that the captain did not tell the sailor to throw the passengers overboard. We hypothesized that participants reading the sailor-focused scenario would attribute less force to the captain, compared to participants reading the captain-focused scenario.

6.2. Results and discussion

As predicted, participants attributed more force to the captain when the scenario focused on the captain (mean: 5.11, standard error: 0.34), versus the sailor (mean: 3.96, standard error: 0.39; $t(51) = 2.22$, $p = 0.031$; Fig. 5). Notably, this difference emerged even though the captain's action and the sailor's action were identical across the two scenarios. The only difference was that one scenario focused more on the captain (the forcer) as the primary agent, and the other scenario focused more the sailor (the forcee) as the primary agent. Increasing focus on the sailor and as a direct result reducing focus on the captain impacted force attributions. When participants' moral focus was on the sailor rather than the captain, they were less inclined to judge that the captain forced the sailor to throw the passengers overboard.

7. General discussion

The current results reveal the impact of moral evaluations on attributions of force. Moral agents are judged as more forceful when they force other agents to act immorally, but, paradoxically, the same agents who are forced to act immorally are also judged as acting more freely and under less force. Directly shifting participants' focus from one moral agent to another, without changing any facts of the narrative, also changed participants' force attributions – the more that participants focused on the forcee (e.g., sailor) instead of the forcer (e.g., captain), the less forceful they judged the forcer to be.

The present findings fit into a larger body of research establishing the impact of moral judgment on judgments

in other domains, including causation (Hitchcock & Knobe, 2009; Knobe & Fraser, 2008), knowledge (Beebe & Buckwalter, 2010), intent (Knobe, 2003), force (Phillips & Knobe, 2009), and counterfactual thinking (Branscombe et al., 1996; N'gbala & Branscombe, 1997). Recent research has also targeted the nature and boundaries of this impact (Guglielmo, Monroe, & Malle, 2009; Machery, 2008; Mallon, 2008). Does this effect reflect a legitimate role for morality in concepts such as force (Knobe, in press; Pettit & Knobe, 2009; Phelan & Sarkissian, 2008; Wright & Bengson, 2009)? Or does it provide evidence for motivated moral reasoning (Adams & Steadman, 2004; Alicke, 2008; Nadelhoffer, in press)? The current paradigm – characterizing the influence of moral evaluations on agent-based evaluations in an interpersonal context – allowed us to test these hypotheses.

7.1. Motivated moral reasoning

Uncovering motivated moral reasoning usually presents a challenge given that standards of evidence for motivated moral reasoning are hard to come by. Directly identifying "moral errors" is difficult given widespread disagreement over the content of "moral truth," how to measure it, and whether it even exists. Previous research has therefore relied on indirect approaches to motivated moral reasoning (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009) by showing, for instance, that many moral judgments represent post hoc rationalizations of emotional biases (Alicke, 2000, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Haidt, 2001; Inbar, Pizarro, Knobe, & Bloom, 2009; Kliemann, Young, Scholz, & Saxe, 2008; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005; Young, Scholz, & Saxe, in press). In one study, for example, participants read about two people who committed incest; participants then made negative moral judgments, driven by their intuitive emotional responses (Haidt, Koller, & Dias, 1993). To rationalize their emotionally mediated moral judgments, participants pointed to the harmful effects of incest even though the scenario stipulated that no harm did or would occur.

The current research remains neutral about the normative status of emotion in moral judgment, but provides another window into motivated moral reasoning. Specifically, we capitalize on the standards of logical consistency as applied to (non-moral) judgments of force. Based on the following features of the findings, we suggest that the influence of moral judgments on force judgments reflects motivated moral reasoning.

First, compared to force judgments made in non-moral contexts, force judgments in moral contexts appear to be inconsistent. Consider a simple case of force: a father forces his child to eat her vegetables. From this, we can correctly infer that the child was forced to eat her vegetables by her father. This example, like the other morally neutral scenarios of the present study, demonstrates people's basic competency for force judgments, but it also contrasts with the paradoxical response pattern observed in the current study: participants judged that X forced Y but that Y was not forced by X (for a morally bad action) and that X did not force Y but that Y was forced by X (for a morally obligatory action).

Second, this logical inconsistency appears to be specific to the moral domain. Participants do not produce this pattern more generally, that is, for physical forces in the natural world (e.g., storms), for agents who bring about neutral (non-moral) outcomes (e.g., throwing cargo overboard), or even for agents who bring about negative outcomes but without the relevant knowledge or intent (e.g., ignorant chief).

Third, participants' force judgments deviated from ordinary force judgments in a systematic way, specifically, to support the assignment of moral blame to the immoral agent in focus. For example, judging that the captain forced the sailor to throw the passengers overboard facilitates blaming the *captain* for his actions, while judging that the sailor was not forced to throw the passengers overboard facilitates blaming the *sailor* for his actions.

Finally, merely manipulating participants' focus and therefore the target of their moral reasoning changed participants' force judgments. Investigating judgments in interpersonal situations allowed us to shift participants' attention from one moral agent to another, while keeping the facts of the narrative constant. This effect may be related to framing effects and order effects which have been argued to reveal errors in moral judgment (Sinnott-Armstrong, Mallon, McCoy, & Hull, 2008).

In sum, participants make logically inconsistent force judgments only for immoral agents to whom they are specifically attending, and in support of their moral judgments of those agents. An account that holds that morality is a core component of the concept of *force* and exerts a legitimate influence on force judgments entails that participants were not making any errors in producing logically inconsistent force judgments for immoral scenarios. By contrast, we have argued that, given participants' logically consistent pattern of responses in non-moral cases, the paradoxical pattern of responses is better understood as the product of motivated moral reasoning.

7.2. Focusing bias

The specific influence of moral focus on force judgments observed in the current study may be related to the more general phenomenon of focusing bias, which has been shown to lead to logical errors in reasoning and decision-making across a number of contexts (Legrenzi et al., 1993). In particular, focusing bias occurs when participants narrowly focus on information provided in the narrative and, as a result, explicitly represent this information in their mental models. These representations usually concern the protagonist, such as events from the protagonist's point of view, and, relevant to the current studies, the protagonist's actions and counterfactual alternatives.

Importantly, the current paradigm featured two protagonists interacting with each other and the environment, in an interpersonal context. This approach allowed us to redirect focus from one protagonist to another, while presenting the same information and, sometimes, identical narratives to participants. We accomplished this by asking about one protagonist versus the other (Experiments 1–4), and also by enhancing or reducing focus on specific protagonists in the text of the narrative (Experiment 5).

The present studies revealed a key consequence of moral focus: logically inconsistent judgments of force, a possible signature of motivated moral reasoning. Meanwhile, prior work has shown that focusing on factors explicitly represented in one's mental models can lead to errors in logic (e.g., modus tollens, Wason's selection task), riskless action choices, and counterfactual reasoning (Legrenzi et al., 1993). These results are therefore consistent with our interpretation that the moral focusing effects observed in the current study reflect a departure from the logical consistency of typical force judgments.

7.3. Counterfactual thinking

The current results also relate to research targeting both the relationship between focusing bias and counterfactual thinking and the relationship between moral judgment and counterfactual thinking. First, when one specific agent is in focus, participants may explicitly represent his action and be more likely to consider counterfactual alternatives to that action – what could the doctor have done differently to avoid the bad outcome? Participants may have then failed to explicitly represent the alternative actions that other agents had, leading to errors in their reasoning. For example, when participants focused on the doctor in the hospital scenario, they may have explicitly represented the doctor's action as leading to the bad outcome, considering primarily the doctor's action and its counterfactual alternatives – what the doctor could have done differently – while ignoring the alternatives available to the chief that may have also prevented the outcome (e.g., not issuing the orders in the first place). This may have led participants to judge that the doctor freely brought about the bad outcome. Meanwhile, when participants focused on the chief, they may have explicitly represented the chief's action as leading to the bad outcome and ignored the alternatives available to the doctor (e.g., refusing to follow the chief's orders), leading them to conclude that the chief forced the doctor. As a result, participants arrived at the logically inconsistent pattern: X forced Y, but Y was not forced by X.

Second, prior work has also shown that participants engage in more counterfactual reasoning for an immoral action (e.g., if only he had not stopped to drink beer on his way home, he could have saved his dying wife) than a moral action (e.g., if only he had not stopped to pick up medication for his parent) (McCloy & Byrne, 2000; N'gbala & Branscombe, 1997). The impact of moral judgment on force judgment might then be mediated by counterfactual reasoning; moral judgment might directly influence counterfactual judgments about whether the agent could have done otherwise. Counterfactual alternatives might then impact judgments of force, since an agent who cannot act otherwise may be perceived as "forced". Thus, when participants focused on the immoral doctor, they might have reasoned: if only he had not followed the chief's orders, the patient would still be alive. The doctor could have done otherwise; the doctor was not forced. By contrast, when participants focused on the immoral chief, they might have reasoned: if only he had not ordered the doctor to kill the patient, the patient would still be alive. As a result, participants might have focused on the chief's causal role in the patient's death, leading to the judgment that the chief forced the doctor to kill the patient.

Additional influences on counterfactual thinking may also help explain the current pattern of force judgments. Previous research has shown that participants are more likely to reason counterfactually in the case of controllable actions performed by agents compared to uncontrollable events (e.g., if only the road had not been blocked) or uncontrollable and unintentional actions (e.g., if only he had not had an asthma attack) (Girotto, Legrenzi, & Rizzo, 1991; Walsh & Byrne, 2007). These differences in counterfactual reasoning track with the differences observed in the current study in force attributions to the captain (more) versus the storm (less), and the intentional chief (more) versus the unintentional chief (less). Future research should directly investigate the precise relationship between counterfactual reasoning and attributions of force.

7.4. Rational alternatives

Are there alternative accounts of the observed paradoxical pattern that allow us to make rational sense of participants' judgments? For instance, if participants had principled reasons for delivering inconsistent force judgments, they might consciously endorse their judgments and not perceive any inconsistency at all. One approach to examining the possibility of rational alternatives is to investigate whether the same logical inconsistency would obtain in a within-subjects design. If participants each made judgments about both the forcer and forcee, would they consciously endorse their conflicting force judgments, or, if provided the opportunity to see the scenarios side by side, would participants self-correct to produce an internally consistent pattern (LeBoeuf & Shafir, 2003; Lombrozo, 2009; Stanovich & West, 1998)?

Previous studies have employed within-subjects designs to investigate the normative status of other framing effects. These studies established robust individual differences: participants with high SAT scores or high Need for Cognition (NC) were more likely to make internally consistent judgments, when two scenarios were presented together in a within-subjects design (LeBoeuf & Shafir, 2003; Stanovich & West, 1998). However, as noted by LeBoeuf & Shafir, 2003, "what high NC respondents are successful at avoiding is inconsistency—not framing per se. Their responses to a second occurrence of a decision problem are likely to be in line with their responses to the first".

Broadly similar to this prior work, pilot data, collected in a within-subjects design using the ship scenario of the present study, revealed individual differences. Overall, the original paradoxical pattern persisted: participants judged that the captain forced the sailor to throw the passengers overboard (mean: 4.84, standard error: 0.20) more than they judged that the sailor was forced to throw the passengers overboard (mean: 4.16, standard error: 0.22; $t(82) = -3.14$, $p = 0.002$). However, this difference was driven by a subset of the 83 participants: 29 subjects showed the predicted difference (mean difference: 2.69, standard error: 0.3), 43 showed no difference, and 11 showed the opposite difference (mean difference: 1.91, standard error: 0.36). The 43 participants who no longer generated different judgments

(compared to the 29 who did) might have engaged in self-correction towards more consistent force judgments once the influence of moral focus was made salient (Legrenzi et al., 1993). Although we did not collect measures of general cognitive ability (e.g., NC), prior work suggests that the individuals who no longer showed the effect, by generating internally consistent judgments, may be higher in NC (LeBoeuf & Shafir, 2003).[6]

Notably, though, the subtle focus manipulation in Experiment 5 (e.g., merely shifting focus without changing the relevant facts of the narrative) is consistent with an unconscious influence of moral evaluation on force judgment, and, more broadly, an "error" account of moral focus as a particular instance of focusing bias (Legrenzi et al., 1993). Future research, however, will be required to determine whether participants consciously or unconsciously modify their force judgments, and to explore individual differences observed within-subjects.

Given that the paradoxical pattern obtained within-subjects for some pilot participants, one could speculate on what principled reasons might be available to govern their judgments. One possibility would be to explain the pattern in the context of causal discounting in the presence of multiple causes (Sloman, 1994). Consider the scenario in Experiment 1 in which the captain orders the sailor to throw the passengers overboard. Participants may see the captain as less forceful, and the sailor as less forced, if they think of the sailor as having an independent reason (aside from the captain's orders) to throw the passengers overboard, namely, for the sailor's own survival. However, we think it is unlikely that causal discounting will be able to account for the full pattern of results across experiments for the following reasons.

First, if the sailor wants to survive himself, it is likely that he wants to survive to the same extent in both the "immoral" and "neutral" scenarios of Experiment 1 – whether the captain orders him to throw passengers overboard (immoral) or cargo (neutral), to save their sinking ship. Yet subjects see the sailor as less forced to act in the immoral scenario.

Second, the immoral nature of the act of throwing passengers overboard might actually be an independent reason for the sailor to *not* want to do so, compared to throwing cargo overboard. Assuming that the sailor is no more eager to throw passengers versus cargo overboard to save his own life, the sailor should not be judged as less forced to throw passengers.[7]

Third, even if participants did judge that the sailor wanted to survive more in the immoral scenario and therefore wanted to throw the passengers overboard more, there's no reason for participants to think that the sailor wanted to survive to differing degrees depending on whether the storm or the captain was described as applying the force, in Experiment 2, or depending on whether the scenario text focused more on the sailor or the captain, in Experiment 5. In both cases, the facts of the narrative are identical, including that the sailor, at the captain's orders, threw passengers overboard to save the sinking ship.

Finally it is doubtful that the results of Experiments 3 and 4, set in the hospital, can be explained by this account. It is unlikely that participants judged the doctor as less forced to kill the patient because of an independent desire to kill the patient; on the contrary, the doctor "really liked the patient and wanted her to recover as quickly as possible". Yet, subjects still judged that the doctor was not forced to prescribe the fatal drug.

Given these reasons, we think it is improbable that participants delivered their force judgments by consciously considering factors such as multiple causes and causal discounting. Instead, we think the current pattern of paradoxical judgments is the product of moral focus. However, future research should investigate the possibility of other consciously accessible principles that could account for the paradoxical pattern of judgments.

8. Conclusions

In sum, the current study suggests that shifting moral focus and therefore the target of motivated moral reasoning – without changing any of the facts – can alter our non-moral evaluations of agents, including whether they were forced to act and whether they forced another to act. As a result, we may find ourselves amidst logical inconsistencies: X forced Y to act immorally, but Y was not forced by X; X did not force Y to act morally, but Y was forced by X. This is the paradox of moral focus.

The present results may also inform social psychological phenomena showing that observers often underestimate the power of situational force and overestimate agents' ability (including their own) to resist such force. Indeed, observers often infer that external behavior (e.g., following evil orders) corresponds to internal dispositions (Gilbert & Malone, 1995), hence our own shock at experimental studies – and real life – when ordinary people knowingly harm others under the right kinds of force (Milgram, 1974; Ross, 1988; Zimbardo, 1973).[8]

[6] Another interesting possibility is that the individuals who continued to show the effect, by generating internally *inconsistent* judgments, are no different (or even higher) in NC. This might suggest that some individuals consciously endorse their own internally inconsistent judgments, because they are sensitive to other explanatory rational principles.

[7] We note though that participants could have inferred that the sailor cares more about survival when he throws passengers overboard from his willingness to incur such a moral cost. Participants could then judge that the sailor had a greater desire to throw the passengers overboard. However, we recognize this as yet another influence of moral judgment on reasoning in another domain – participants' moral judgments may very well influence both attributions of force as well as attributions of intent and desire (Knobe, in press).

[8] Would greater sympathy for the forcee (e.g., the average participant in the Milgram experiment) reduce both attributions of moral blame and attributions of freedom to do harm? A corollary prediction of our account is that upon watching a video of the Milgram experiment, observers might mitigate blame to the extent that they observed the abundant visual cues of participants' suffering while being forced to follow immoral orders. Sympathy for the participants should attenuate the amount of moral blame and therefore reduce judgments of freedom. In other words, to the extent that we do not want to blame the Milgram participants, we also want to say that they were truly forced to administer the shocks. This pattern would complement the present finding that harsh moral judgments enhance attributions of freedom (to do harm).

Demanding an evaluation of – or otherwise focusing attention on – either the agent issuing the orders or the agent following them may be akin to placing one defendant versus the other on trial in a court of law. Like our participants, jurors may establish legal elements of duress and coercion (i.e. force judgments) differently, depending on who is on trial, and for what crime. When focused on the forcer as the immoral agent, people may be motivated to attribute more force, reasoning, consciously or unconsciously, that forcing another to do harm requires great force. When focused on the forcee as the immoral agent, people may be motivated to attribute more freedom, reasoning that one must be able to resist authority in such cases. These differences may result in logical inconsistencies as observed in the current study. How these findings relate to the role of moral focus in other legally and morally relevant aspects of folk psychology is a challenge for future research.

Acknowledgments

We gratefully acknowledge Rebecca Saxe, Jorie Koster-Moeller, Laura Schulz, Fiery Cushman, Chris Baker, Edouard Machery, David Rose, Kurt Gray, Dylan Murray, Mark Alicke, and especially Joshua Knobe for helpful discussion and comments.

References

Adams, F., & Steadman, A. (2004). Intentional action and moral considerations: Still pragmatic. *Analysis, 64*, 268–276.

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556–574.

Alicke, M. D. (2008). Blaming badly. *Journal of Cognition and Culture, 8*(1–2), 179–186.

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin, 34*(10), 1371–1381.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership and men*. Pittsburgh: Carnegie Press.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*, 74–85.

Beebe, J., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language, 25*, 474–498.

Borg, J. S., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An FMRI investigation. *Journal of Cognitive Neuroscience, 18*(5), 803–817.

Branscombe, N., Owen, S., Garstka, T., & Coleman, J. (1996). Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome? *Journal of Applied Social Psychology, 26*, 1042–1067.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition, 108*(2), 353–380.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PLoS ONE, 4*(8), e6699.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108*(1), 281–289.

Darley, J. M., & Shultz, T. R. (1990). Moral rules – Their content and acquisition. *Annual Review of Psychology, 41*, 525–556.

Frosch, C., Johnson-Laird, P., & Cowley, M. (2007). *It's not my fault, your honor, i'm only the enabler*. Paper Presented at the Proceedings of the 29th Annual Cognitive Science Society, Austin, TX.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*(1), 21–38.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica, 78*, 111–133.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*(3), 505–520.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Guglielmo, S., Monroe, A., & Malle, B. (2009). At the heart of morality lies folk psychology. *Inquiry, 52*, 449–466.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*(4), 613–628.

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

Harvey, J., Harris, B., & Barnes, R. (1975). Actor–observer differences in the perceptions of responsibility and freedom. *Journal of Personality and Social Psychology, 32*(1), 22–28.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy, 11*, 587–612.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion, 9*(3), 435–439.

Jones, E., & Harris, V. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*, 1–24.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*(12), 2949–2957.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–193.

Knobe, J. (in press). Person as scientist, person as moralist. *Behavioral and Brain Sciences*.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (pp. 441–444). Cambridge, MA: MIT Press.

LeBoeuf, R., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making, 16*, 77–92.

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (1993). Focussing in reasoning and decision making. *Cognition, 49*(1–2), 37–66.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science, 33*, 273–286.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language, 23*, 165–189.

Malle, B. (2006). The relation between judgments of intentionality and morality. *Journal of Cognition and Culture, 6*, 61–86.

Mallon, R. (2008). Knobe versus Machery: Testing the trade-off hypothesis. *Mind & Language, 23*(2), 247–255.

McCloy, R., & Byrne, R. (2000). Counterfactual thinking about controllable actions. *Memory & Cognition, 28*, 1071–1078.

Mikhail, J. M. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal Psychology, 67*, 371–378.

Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row Publishers, Inc.

N'gbala, A., & Branscombe, N. (1997). When does action elicit more regret than inaction and is counterfactual mutation the mediator of this effect? *Journal of Experimental Social Psychology, 33*, 324–343.

Nadelhoffer, T. (in press). Intentional action and intending: Recent empirical studies. *Philosophical Psychology*.

Nario-Redmond, M., & Branscombe, N. (1996). It could have been better or it might have been worse: Implications for blame assignment in rape cases. *Basic and Applied Social Psychology, 18*, 347–366.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*(4), 663–685.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language, 24*(5), 586–604.

Phelan, M., & Sarkissian, H. (2008). The folk strike back: Or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies, 138*(2), 291–298.

Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry, 20*(1), 30–36.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin, 121*(1), 133–148.

Ross, L. (1988). Situationist perspectives on the obedience experiments. *Contemporary Psychology, 33*, 101–104.

Sinnott-Armstrong, W., Mallon, R., McCoy, T., & Hull, J. (2008). Intention, temporal order, and moral judgments. *Mind & Language, 23*(1), 90–106.

Sloman, S. (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition, 52*, 1–21.

Snyder, M., & Jones, E. (1974). Attitude attribution when behavior is constrained. *Journal of Experimental Social Psychology, 10*, 585–600.

Stanovich, K., & West, R. (1998). Individual differences in framing and conjunction effects. *Thinking and Reasoning, 4*, 289–317.

Uhlmann, E., Pizarro, D., Tannenbaum, D., & Ditto, P. (2009). The motivated use of moral principles. *Judgment and Decision Making, 4*, 479–491.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476.

Walsh, C., & Byrne, R. (2007). How people think "If only..." about reasons for actions. *Thinking and Reasoning, 13*, 461–483.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*(8), 383–388.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*(10), 780–784.

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*(2), 283–301.

Wright, J., & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind & Language, 24*(1), 24–50.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology, 1*, 333–349.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47*(10), 2065–2072.

Young, L., Scholz, J., & Saxe, R. (in press). Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts. *Social Neuroscience*.

Zimbardo, P. G. (1973). On the ethics of intervention in human psychological research: With special reference to the Stanford Prison Experiment. *Cognition, 2*(2), 243–256.