

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Psychology

2014

Intentional Harms Are Worse, Even When They're Not

Ames, Daniel Lukas

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01m326m1878>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

INTENTIONAL HARMS ARE WORSE, EVEN WHEN THEY'RE NOT

DANIEL LUKAS AMES

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

PSYCHOLOGY

ADVISOR: SUSAN T. FISKE

JUNE 2014

© Copyright by Daniel Lukas Ames, 2014. All rights reserved.

Abstract

This dissertation identifies a tendency for observers to view intentional harms as larger in magnitude than they in fact are. Chapter 1 (Ames & Fiske, 2013a) provides the first demonstrations of this bias, and provides preliminary evidence concerning mechanism. Chapter 2 (Ames & Fiske, under invited revision) extends this work into the realm of motivated social cognition. Specifically, this chapter highlights the apparent failure of well-characterized mental inference (*mentalizing*) systems to prevent the magnification of intentional emotional harms. Chapter 3 (Ames & Fiske, in preparation) describes a series of experiments designed to provide insight into the psychological mechanisms that underlie the magnification of intentional harms. A final section summarizes the work to date and suggests potential legal applications and theoretical considerations.

TABLE OF CONTENTS

Abstract.....	iii
Acknowledgements.....	v
Chapter 1: The magnification of intentional harms.....	1
Chapter 2: Motivated mentalizing: Oversensitivity to intentions leads to magnified emotional harm perception.....	18
Chapter 3: But is it really motivated?	57
Summary and applications.....	73
References.....	79
Appendices.....	94

Acknowledgements

The best thanks that I can offer to my dissertation committee is to keep my acknowledgements section as brief as possible. My dissertation committee won't be receiving the best thanks that I can offer them, but I thank them nonetheless: Susan Fiske, Alex Todorov, Alin Coman, Joel Cooper, and Nick Turk-Browne.

Susan Fiske has been an exemplary advisor. She is every bit as warm and competent as this joke is worn and commonplace; and if she's tired of hearing it—if she's tired of her students writing embarrassingly long paragraphs about her in their dissertation acknowledgements sections—paragraphs that she knows the other committee members are reading at this very moment—well, she has only her own attributes to blame. Her students of years past have more or less exhausted the world's supply of positive adjectives that one can (reasonably) ascribe to one's advisor. I refuse to be unoriginal by re-appropriating something they've already written. Instead, I'll be unoriginal by re-appropriating something she's already written: She is a Fixture in the Field who refuses to remain fixed. She doesn't spend much time defending her intellectual turf. That's not to say that she has taught me to ignore criticism, or that a sustained intellectual debate can't be useful and rewarding; I just think she's too interested in what's over the next horizon to spend much time looking back. This and her work ethic are the two things that I most admire about her. The passive voice, however, is something about which she and I will have to disagree.

After thanking one's committee members and, particularly one's primary advisor, it is customary to thank one's parents. I'd like to think that I'd have the sense of gratitude and decency to thank them even if it weren't customary. Though in that counterfactual world, my sense of gratitude and decency would just be one more thing for which I'd have to thank them. Best to avoid counterfactuals then. Thank you Mom and Dad. I'm exceptionally lucky to have you as

parents. I know how important it was to you that Ben and I get a good education; and I can at least guess at some of the sacrifices that you might have made in order to make that happen for us. Four years of college tuition can buy a *lot* of gardening gloves and windsurfer boards! I hope that in at least some sense, this dissertation is a tangible signal to you that those gardening gloves and windsurfer boards weren't sacrificed completely in vain. I hope you don't ever subject yourself to actually reading this tangible signal though. It's really long.

CHAPTER 1: THE MAGIFICATION OF INTENTIONAL HARMS ¹

Chapter Summary

People and societies seek to combat harmful events. However, because resources are limited, every wrong righted leaves another wrong left unchecked. Responses must therefore be calibrated to the magnitude of the harm. One underappreciated factor that affects this calibration may be people's oversensitivity to intent. This chapter reports a series of studies in which participants saw intended harms as worse than unintended harms, even though the two harms were identical. This harm-magnification effect occurred for both subjective and monetary estimates of harm, and it remained when participants were given incentives to be accurate. The effect was fully mediated by blame motivation. These findings suggest a novel explanation as to why laypersons and policymakers may focus attention and resources on intentional harms to the neglect of unintentional (but equally damaging) harms.

¹ The work described in this chapter has been published as Ames, D.L. & Fiske, S. T. (2013a). Intentional harms are worse, even when they're not. *Psychological Science*, 24(9), 1755–1762.

Introduction

Motivations induce biases that help people see what they want to see (Kunda, 1990). This process can be described via an analogy with adversarial legal systems, in which lawyers (like motivational forces) prefer one verdict over another. They therefore restructure or even distort the facts in a way that tips the judge (or judgment processes) toward the desired conclusion. However, even as motivated decision makers act as lawyers, they tend to see themselves as impartial judges (Ditto, Pizarro, & Tannenbaum, 2009). This “illusion of objectivity” (Pyszczynski & Greenberg, 1987) allows motivated reasoning to subtly affect people’s genuine beliefs.

Judgments about moral and immoral acts may be especially subject to such motivational biases (Ditto et al., 2009). Most moral judgments are emotional, complex, and subjective—precisely the conditions under which the effects of motivated reasoning tend to be largest (Ditto & Boardman, 1995; Dunning, Leuenberger, & Sherman, 1995). However, some motivations may also be sufficiently powerful to influence the perception of straightforward, objective information. In this chapter, we examine whether moral motivations might cause people to see certain kinds of harms—specifically, intentional harms—as worse than they actually are.

Studies on belief in a just world (Lerner, 1965), defensive attributions (Shaver, 1970), retributive justice (Darley & Pittman, 2003), criminal responsibility (Gebotys & Dasgupta, 1987), and moral psychology (Gray & Wegner, 2010; Haidt & Kesebir, 2010; Knobe et al., 2012) all converge to show that when people detect harm, they become motivated to blame someone for that harm. This demonstrably powerful motivation has been variously indexed as the need to assign blame, to express moral condemnation, and to dole out punishment (Alicke, 2000; Haidt,

2001; Robinson & Darley, 1995). Often, these three needs co-occur, and they may represent different, but related, components of humans' moral response to harm. For convenience, we refer to these components collectively as *blame motivation*, and we suggest that people seek (though not necessarily consciously) to satisfy this motivation when confronted with harm.

If people (acting as motivated lawyers) want to blame harm-doers (Alicke & Davis, 1990), what evidence could they emphasize (or distort) to make their case more compelling? We suggest that one tactic would be to imply that harm-doers caused more harm than they actually did. That is, in order to satisfy blame motivation, one's inner lawyer might marshal more evidence for harm than actually exists. From this line of reasoning, we argue that harmful acts that lead to especially large amounts of blame motivation might also lead to exaggerated perceptions of harm. But what kinds of harmful acts are especially likely to induce blame motivation?

Though we suspect that this question has several answers, numerous behavioral (and, recently, neuroscientific) experiments have demonstrated that *intentional* harms make people want to blame, condemn, and punish more than unintentional harms do (Alicke, 1992; Darley & Pittman, 2003; Young & Saxe, 2009). For example, people are approximately twice as likely to punish someone who intentionally cheated (by not reciprocating in an economic game) than to punish someone who made an innocent mistake (because of inexperience with the game; Hoffman, McCabe, & Smith, 1998). People are notoriously sensitive to harmful intentions (Gollwitzer & Rothmund, 2009), and even exposure to fictional characters with harmful intentions can change subsequent trust behavior in real life (Rothmund, Gollwitzer, & Klimmt, 2011). If, as we propose, intent leads to blame motivation, and blame motivation, in turn, leads to increased perceived harm, it follows that harms that are committed intentionally should be seen

as worse than they actually are, and as worse than identical harms committed unintentionally.

This is the main hypothesis we aimed to test in the present research. We further predicted that the effect of intentionality on perceived harm would be mediated by blame motivation. In Studies 1 through 3, we presented participants with simple vignettes that described harmful events and manipulated whether those harms were framed as intentional or unintentional (while holding constant the actual harms). Participants then judged how much harm befell the victims. In Studies 4 and 5, we tested whether this bias would remain even when objective, accurate information about the amount of harm was readily available, and even when participants were given incentives to be accurate.

Participants

Participants ($N = 675$) were recruited using Amazon Mechanical Turk and then redirected to an experimental website. They were paid standard rates ranging from 20¢ to \$2.00 for participating in these studies. Recruitment was limited to the United States, and to individuals with at least a 95% approval rating for their previous work on MTurk. Current research suggests that samples recruited via Mechanical Turk are more demographically and cognitively representative of national distributions than are traditional student samples, and that data reliability is at least as good as that obtained via traditional sampling (e.g., Buhrmester, Kwang, & Gosling, 2011; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010). Participants who had completed any previous study in this line of work were automatically prevented from accessing the recruitment page. Of the 675 who were recruited, 39 participants failed to complete the task. Of the remaining 636 participants, 14% were excluded for failing the manipulation check or basic checks for meaningful responses. For complete information on these checks and the number of persons excluded by each one, see Participant Exclusion Criteria in Appendix A.

Study 1

Method

Participants ($N = 80$) in our first study read a vignette about a CEO of a small company where employees' compensation was partially based on company profits (i.e., profit sharing). The CEO made an investment that performed poorly and therefore cost his employees money. Participants were randomly assigned to either the *intentional-harm* condition ($N = 39$) or the *unintentional-harm* condition ($N = 41$), which differed in whether the vignette specified that the CEO harmed his employees intentionally (he thought that “they would work harder to increase profits in the future if they made less money now”) or unintentionally (he thought the investment was a good one that would make money, but “the investment failed”). (See Appendix B for the full texts used in both conditions.)

Although this framing explaining the CEO's intents varied across conditions, the actual harms that participants were asked to judge were identical:

The employees made less money than they otherwise would have. This caused them some anxiety and made paying off credit cards and taking family vacations harder for them. Two employees saw a minor dip in their credit scores because of late payments—though nobody suffered truly dire financial hardship or came close to losing their home or their car or anything.

To avoid the confound that the employees would have felt more *victimized* by intentional harm than unintentional harm (which would likely have made intentional harm *actually* worse; Gray &

Wegner, 2008), we stipulated that the employees in both conditions were unaware of the investment. The company made less money, and therefore its profitsharing employees also made less money; however the employees had no reason to suspect foul play, incompetence, or any other cause more in one condition than in the other. Thus, the harms to the employees were truly identical (both financially and psychologically) in the two conditions.

Nonetheless, we predicted that the harms would be judged as different. Because our hypotheses concerned harm in general, we began by asking participants about harm in general (specific, objectively quantifiable harms were examined in Studies 4 and 5). Participants were asked: “How much harm would you say Terrance’s [the CEO’s] investment in the software development project did to the employees?” Responses were made on a sliding scale ranging from 0 (*none*) to 100 (*the most harm Terrance’s actions as CEO could possibly cause them*). Thus, all participants judged the same act (the investment) by the same person (the CEO) and the same resulting harm to the same people (the employees). Nonetheless, we predicted that participants would see the harms as greater when framed as intentional rather than unintentional, and that this effect would be due to (fully mediated by) differences in blame motivation. Following our earlier definition of blame motivation (desire to blame, morally condemn, and punish), we assessed this construct via a three-item ($\alpha = .86$) index (sliding scale from 0 to 100): “To what extent do you think Terrance deserves blame for making the investment?”; “To what extent do you think Terrance should be morally condemned for making the investment?”; and “To what extent would you want Terrance to be punished for making the investment?”

Results

Intent. A manipulation check confirmed that participants indeed saw the harm as more intentional in the intentional-harm condition ($M = 65.13$, $SD = 36.65$) than in the unintentional-

harm condition ($M = 4.20$, $SD = 9.81$), $p < .001$, $d = 2.27$ (scale from 0 through 100).

Harm. Our key prediction was confirmed. Although the harms to the employees were described identically in the two conditions, participants thought that the employees were more harmed in the intentional-harm condition ($M = 65.95$, $SD = 23.20$) than in the unintentional-harm condition ($M = 47.59$, $SD = 26.40$), $p = .001$, $d = 0.74$ (Fig. 1).

Blame motivation. As predicted, participants were more motivated to blame Terrance when he had harmed the employees intentionally rather than unintentionally, $t(78) = 8.74$, $p < .001$. (The difference between conditions was also significant when ratings for each blame motivation item were examined separately, $ps < .001$.) Moreover, blame motivation fully mediated the effect of perceived intentionality on magnitude of perceived harm, 10,000 bootstrapped samples, 95% CI

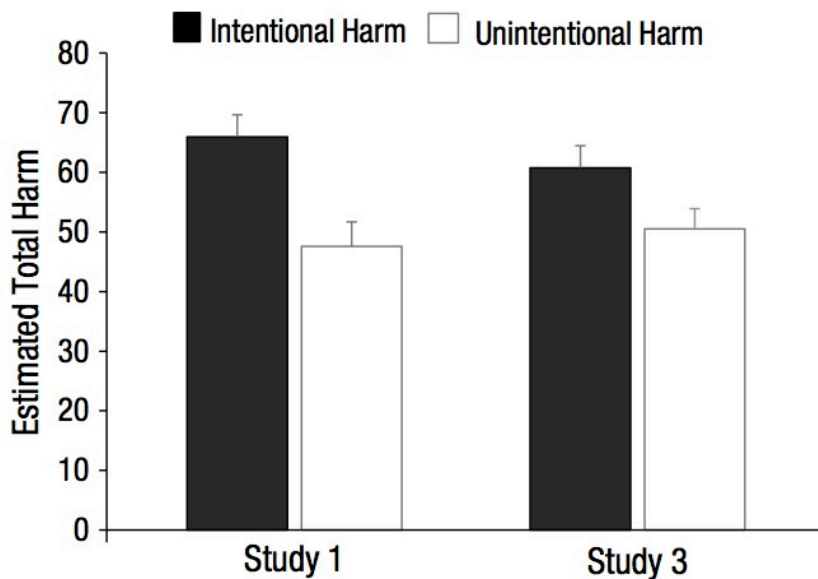


Fig. 1. Results from Studies 1 and 3: perceived harm to employees (scale from 0 through 100) in the intentional- and unintentional-harm conditions. Error bars represent standard errors.

[.43, .78] completely standardized, Sobel $Z = 5.00$, $SE = 0.08$, $p < .001$. In contrast, the alternative model, with perceived harm leading to blame motivation rather than vice versa, was a worse description of the data, and a bootstrapping analysis

fully corroborated these findings (see Appendix C for supplementary mediation analyses).

Discussion

Study 1 provides initial support for the hypothesis that people see intended harms as worse than (identical) unintended harms. But why? Our theoretical account suggests that intent provoked blame motivation and that, because of their motivation to build a case against the CEO, participants exaggerated how much harm had been done. Mediation analyses supported this model; however, an alternative possibility exists. Legal scholars point out that CEOs who act criminally or immorally not only harm individuals but also create broader social harms, to which people are sensitive (e.g., Cohen, 1988). From this perspective, the CEO's actions in this study would have been perceived as a betrayal of public trust—a harm against the fabric of society in general (Darley & Huff, 1990). This perceived harm to society in general may have leaked into participants' ratings of the harms to a circumscribed set of specific employees.

However, it should not be assumed that participants held this view about harm to society, or that they were incapable of separating harms to society from harms to specific individuals. A conservative test as to whether different judgments of harm to society could have produced the effects in Study 1 would be to test whether such differences in judgment exist at all. If not, then it would be difficult to argue that they explain the results.

Study 2

We therefore reproduced Study 1 with only one change: Instead of asking participants how much harm Terrance's investment in the software development project did to the employees, we asked, "How much harm would you say Terrance's investment in the software development project did to society in general?"

Although intentionality was again manipulated effectively (intentional harm: $N = 46$, $M = 71.70$, $SD = 34.73$; unintentional harm: $N = 47$, $M = 3.85$, $SD = 10.19$, $p < .001$), and the blame-

motivation scale remained reliable ($\alpha = .81$), there was no magnification of harm—that is, no effect of intent on perceived harm to society (intentional harm: $M = 45.63$, $SD = 24.51$; unintentional harm: $M = 38.40$, $SD = 28.34$), $p = .19$.

It therefore appears unlikely that different perceptions of harm to society explain the effects observed in Study 1, as no such differences exist. However, it could be argued that Study 2 might simply constitute a failure to replicate. Yet another possibility is that participants in Study 1 (consciously or unconsciously) accounted for future harms in the intentional-harm condition, reasoning that the CEO who intentionally harmed his employees once (and did not get caught) might do so again, thus causing more harm. We conducted Study 3 to test these two possibilities, directly replicating the intentional and unintentional harm conditions from Study 1 (to address the “failure to replicate” alternative hypothesis) and adding a third condition in which future harm was not a factor (to address the “future harms” alternative hypothesis).

Study 3

Procedures were identical to those in Study 1 for participants in the intentional-harm ($N = 41$) and unintentional harm ($N = 38$) conditions. All effects observed in Study 1 replicated. The intent-magnifies-harm effect was again significant, $p = .04$, $d = 0.48$, and the mean harm ratings in both conditions were statistically identical to those observed in Study 1—intentional harm: $M = 61.26$, $SD = 23.67$, $p = .29$; unintentional harm: $M = 50.50$, $SD = 21.18$, $p = .59$ (see Fig. 1). Participants also expressed more blame motivation in the intentional-harm condition than in the unintentional-harm condition, $t(77) = 8.26$, $p < .001$. The blame-motivation index was again reliable ($\alpha = .91$) and again fully mediated the effect of intentionality on perceived harm, 95% CI [.08 .30], completely standardized, Sobel $Z = 3.19$, $SE = 0.04$, $p = .001$. Bootstrapping analyses

again corroborated these results, and the results of both mediational approaches supported the predicted model (intent → blame motivation → harm) over the alternative model (intent → harm → blame motivation; see Appendix C for supplementary mediation analyses).

As noted, to test whether participants' harm ratings might have reflected anticipated future harms, we also ran a third condition: the *intentional-harm/fired* condition ($N = 42$). This was identical to the intentional-harm condition except that the vignette explicitly precluded the possibility of the CEO perpetrating additional harms against the employees by stipulating that he was caught and fired for his actions (see Appendix B for full vignette text). This change, however, made no difference in perceived harm (intentional harm: $M = 61.05$; intentional harm/fired: $M = 61.57$), $p = .96$. This pattern of results suggests that the observed harm-magnification effect was not an artifact of anticipated future harms to the employees.

Studies 1–3: Summary and Limitations

Studies 1 through 3 provided initial evidence for the intent-magnifies-harm effect and its proposed mechanism, while ruling out some possible confounds. However, they also left at least two questions unresolved.

First, does this effect hold for harms that are objectively quantifiable and ecologically meaningful? Because our hypotheses concerned harm in general, it seemed appropriate to begin by asking participants “how much harm” they perceived without forcing upon them any specific definition of harm. However, it is unclear what “61.5 units of harm” (for example) means (Blanton & Jaccard, 2006). Second, what is the direction of the effect? Our theoretical account predicts an exaggeration of intentional harms. However, Studies 1 through 3 could not test directionality, as there was no “ground truth” as to how much harm actually occurred. In Studies

4 and 5, we aimed to address these questions.

We also sought to generalize our findings beyond comparisons of two human harm-doers. In Studies 4 and 5, we compared a disaster caused by human intent with the same disaster caused by nature (no intent). This comparison is particularly relevant to ongoing debates concerning the allocation of government funds to mitigate disasters that feel highly intentional (e.g., terrorism) relative to those that do not (see the General Discussion at the conclusion of this chapter).

Study 4

Method

Participants read a vignette that described a group of people suffering from a water shortage. The proximal cause of the shortage (a specific river drying up) and all of the consequences (crops lost, etc.) were identical across conditions, but the conditions (randomly assigned) varied in how the cause of the damages was framed (see Appendix B for full vignette text). Participants in the intentional-harm condition ($N = 29$) read that the river dried up “because a man who lived in a nearby town diverted the flow” with the intent to cause the harm that occurred, whereas participants in the unintentional-harm condition ($N = 26$) read that the shortage was naturally occurring (“due to a lack of rain”).

After reading the vignette, participants were instructed that they would see all of the harms caused by the drought (in dollars) and then estimate how much total harm had occurred (in dollars). They were warned that the harms would be presented quickly, and that they should pay close attention. Next, they saw seven harms that resulted from the water shortage (e.g., “medical supplies used: \$80.00”). These harms appeared for 2,000 ms each, separated by 300 ms, and in a fixed order. Immediately after viewing these numbers, participants were asked to estimate their

sum (response time limit: 10 seconds). The list of harms was identical in the two conditions.

Results

Participants in the unintentional-harm condition estimated the drought damage accurately; their average estimate of \$2,753 ($SD = 4,518.28$) was close to the correct answer of \$2,862, $t(25) = 0.40$, $p = .70$. By comparison, participants who thought that those same harms were intentional overestimated their sum by a substantial margin ($M = \$5,120$, $SD = 1,396.32$), $t(28) = 2.69$, $p = .01$, $d = 0.50$ (see Fig. 2). The two estimates were also significantly different from one another, $t(33.85) = 2.68$, $p = .01$, $d = 0.71$ (corrected for unequal variances). These results confirm the hypothesized direction of the effect, showing that intentional harms were magnified, whereas unintentional harms were assessed accurately. Moreover, the effect in dollars (about \$5,100 vs. about \$2,800) is more interpretable than the effects observed using the arbitrary metric in Studies 1 through 3.

The magnitude of this effect is perhaps surprising given the apparent straightforwardness of the task (i.e., estimating the sum of seven numbers). Results from the unintentional-harm condition suggest that people are capable of estimating the listed harms accurately (at least on average), so why did they grossly overestimate the harms in the intentional-harm condition? Perhaps participants misunderstood the instructions and added punitive damages or fines to their estimates. Or perhaps they understood the instructions but, in the absence of any incentive for accuracy, chose to express their blame motivation through a deliberately inflated estimate. To test these possibilities, we conducted a fifth study, in which we replicated the procedure of Study 4 except that we provided financial and competitive incentives for accuracy and also included a third condition in which participants had to explicitly separate punitive damages from their estimates of the objective harms.

Study 5

The materials and procedure for the intentional-harm and unintentional-harm conditions in Study 5 were identical to those in Study 4, except that participants were told that the 5% of participants who made the most accurate estimates would have their pay for participation multiplied by 10. Again, participants in the unintentional-harm condition ($N = 65$) estimated the damage accurately ($M = \$2,773$, $SD = 1,932.66$), $t(64) = 0.38$, $p = .71$, but participants in the intentional-harm condition ($N = 71$) overestimated the damage ($M = \$4,176$, $SD = 4,194.62$), $t(70) = 2.64$, $p = .01$, $d = 0.31$ (see Fig. 2). Also as in Study 4, the estimates of the two groups were significantly different from one another, $t(100.38) = 2.54$, $p = .01$, $d = 0.43$ (corrected for unequal variances). Although the magnitude of the bias in the intentional-harm condition was somewhat smaller when people were paid to be accurate than when they were given no such incentive, neither of the two groups' estimates in Study 5 differed significantly from the corresponding estimate in Study 4—intentional-harm condition: $p = .32$; unintentional-harm condition: $p = .96$. Thus, even with financial and competitive incentives to be accurate, participants still saw intentional harms as larger than they actually were.

Finally, 61 participants completed the *intentional harm/separate-estimates* condition. These participants were presented with the same vignette used in the intentional-harm condition. However, instead of promoting accuracy by using monetary incentives, we sought to promote accuracy by eliminating any room for discretionary adjustments of the estimate. Participants were asked to provide two distinct monetary values: their estimate of the objective sum of the harms and their punitive assessment (counterbalanced order). In one box, we asked them to “estimate the total of the damages, in dollars (the sum of the numbers you just saw).” In a

separate box, they were asked to indicate “how much extra money should that person have to pay in fines/penalties/ compensation beyond the actual costs as shown (that is, beyond the sum of the numbers you just saw)?” Participants had 12 seconds to make each response. The estimates of objective harm ($M = \$3,627$, $SD = 2,001.90$) were still significantly larger than the correct answer, $t(60) = 2.98$, $p = .004$, $d = 0.38$, and were similar ($p = .35$) to those of the incentivized participants who did not receive these explicit instructions. Moreover, these biased harm estimates were positively correlated with punitive assessments, $r(54) = .39$, $p = .003$, which provides further evidence that biases in participants’ objective estimates were not due to some participants “moving” desired punitive damages into their estimates of objective damages. In sum, although we suspect that the very large effect observed in Study 4 may have been artificially inflated by the inclusion of punitive damages in harm ratings, it seems unlikely that the same concern applies to Study 5.

General Discussion

This chapter reports five studies suggesting that people overestimate the magnitude of intentional harms. Studies 1 through 3, which allowed participants to apply their own definition of “harm,” show that the intent-magnifies-harm effect emerges when people judge identical actions that are committed by identical people and result in identical harms to identical victims. They also indicate that the effect is not an artifact of outrage about harm to society in general or of anticipated future harms. Studies 1 and 3 show that the effect is reliably mediated by blame motivation.² Studies 4 and 5 demonstrate the directionality of the effect and its durability in the

² We hasten to add that statistical mediation should always be taken as a starting point for establishing psychological causality. We return to this issue in Chapter 3, which reports experiments that more directly address the mechanisms underlying these effects..

face of incentives to be accurate.

These findings relate to a number of ongoing lines of research. Intent is among the most important theoretical constructs for “blame experts,” that is, psychologists, philosophers, and legal scholars concerned with blame and moral judgment (Borgida & Fiske, 2008; Fiske, 1989). At a mechanistic level, the present work comports well with Alicke’s (2000) culpable-control model of blame assessment, which prioritizes motivational factors in shaping the evaluation of evidence. At the same time, this work suggests patterns of causality that are different from those demonstrated by the culpable-control model, showing that blame motivation can shape the very perceptions of harm that were supposed to have caused that motivation. Similarly, whereas the Knobe effect (Knobe, 2003) demonstrates that people are more likely to see harmful than helpful acts as intentional, the present work demonstrates that people are likely to see intentional acts as more harmful than unintentional ones.

The present work also aligns well with previous research on distortions of long-term memory. For example, in an experiment by Pizarro, Laney, Morris, and Loftus (2006), participants learned about a man who had walked out on a restaurant bill. The researchers then presented several pieces of postevent information (Loftus, 1997), to influence recall for the event. Specifically, they described the restaurant patron either as a good, caring person acting under extraordinary personal circumstances (he received a phone call saying that his daughter was ill, had to rush to the hospital, and paid the bill later, calling from the hospital to apologize profusely) or as a morally loathsome individual who walked out on as many restaurant bills as possible for fun, despite having plenty of money, and did so after mistreating the waitstaff and other restaurant patrons. One week later, when participants were given a surprise memory test, those who had been given negative moral information misremembered some of the items on the

unpaid restaurant bill as being more expensive than they had actually been. Interpreting the specific impact of moral information in this study is complicated by the fact that an identical effect emerged in the control condition, in which no information on moral character was given. Nonetheless, this result is broadly consistent with many previous studies showing that suggestive postevent information can cause people to misremember events over time (Loftus, 1997). The present research extends this idea, showing that neither postevent information nor lengthy windows for forgetting seem to be necessary to produce such distortions, while also providing preliminary evidence for an underlying mechanism and for robustness in the face of accuracy incentives.

Given that the phenomenon under examination makes people less accurate in their judgments, how might it be psychologically or socially functional? First (as we have said), seeing more harm than is actually present can be psychologically functional in helping people to satisfy the motivation to blame. Second, because people live in hyper-interdependent societies, and have from their evolutionary beginnings (Kurzban & Neuberg, 2005), it is socially functional for them to be extremely sensitive to individuals intentionally harming one another (Hoffman et al., 1998). This oversensitivity to intentional harms helps protect social groups from malevolent individuals, but may also cause individuals to overestimate the magnitudes of intentional harms.

One policy implication arising from this work is that perceived intent may inflate legal penalties for property damage. This suggestion, which has implications for tort law and insurance regulations, is supported by “an unpredicted and potentially interesting effect” observed by Darley and Huff (1990, p. 181). They found that deliberate physical damages to property were judged as more expensive than accidental damages. The authors noted, however,

that because this comparison was not the original goal of the study in which the effect was found, the experimental design left room for multiple interpretations (including that damage assessments were mixed with punitive damages). Also, in the only study designed to test this comparison specifically, the result was not statistically significant. Still, the findings are potentially important, because they point to a discrepancy between the way humans beings actually reason and the underlying assumptions of the legal system.

Our studies may also have implications for understanding policy judgments regarding government budgets. Because nations have finite resources, every wrong righted leaves another wrong left unchecked. It has been argued (Gilbert, 2011) that resources are overallocated to harms that feel highly intentional (e.g., terrorism), even in the face of data suggesting that humanitarian interests might be better served by dedicating some of these resources to combating other harms (e.g., global warming). The present work suggests a novel psychological mechanism for this phenomenon: Intentional harms receive more funding and attention not only because of political imperatives and moral reactionism (Gilbert, 2011; McGraw, Todorov, & Kunreuther, 2011), but also because intent magnifies the perceived harms themselves.

Of course, there are many reasons why political power holders might invest more heavily in fighting terrorism than in other kinds of interventions. Some of these reasons are likely very sensible, and some are likely less so. Separating these two kinds of reasons is important for calibrating responses to global events. The present work suggests that one underappreciated factor currently affecting this calibration may be people's oversensitivity to intentional harms, which leads them to overestimate the magnitude of those harms (perhaps even when they are incentivized to be accurate). Given the dollars and humanitarian issues at stake, this possibility may be worth additional investigation.

CHAPTER 2: MOTIVATED MENTALIZING: OVERSENSITIVITY TO INTENTIONS LEADS TO MAGNIFIED EMOTIONAL HARM PERCEPTION³

Chapter Summary

The previous chapter discussed some possible legal and political implications of intentional harm magnification. Here we examine the implications of this effect in an entirely different domain: the inference of transient mental states (*mentalizing*), and specifically, inferring how another person is likely to feel in a given situation (*affective mentalizing*). Affective mentalizing is widely studied; however the question of how such mentalizing is affected by human *motivation* remains largely ignored. When humans mentalize, they are not always disinterested social observers. Rather, as people engage in everyday mindreading, they sometimes *prefer* some interpretations of others' minds over others.

This is a broad and largely unexplored idea, likely worthy of empirical investigation beyond the purview of the present work. The findings observed in Chapter 1 suggest one possible starting place for such work: *blame motivation* may be one kind of motivation that shapes social perception, and conceptualizing hurt feelings as *emotional harm* allows us to apply the findings discussed in Chapter 1 to make a prediction in this new domain: just as blame motivation makes people hypersensitive to *financial* harms (making them seem larger than they are), blame motivation may likewise make people oversensitive to *emotional* harms (making victims seem more emotionally hurt than they really are). This chapter provides initial evidence for this idea, with blame motivation consistently mediating the predicted effect. This effect is

³ The work described in this chapter is currently under invited revision for publication in *The Journal of Experimental Social Psychology*.

observed across varied paradigms and stimuli, including a modified classic false belief task (Study 1), a short story emphasizing a single protagonist's inner mental life (Study 2), and a humorous video (Study 3). These findings appear to be in tension with a well-established literature on mental state reasoning processes. However we argue that this tension is an argument for expanding, rather than abandoning, existing theories.

Introduction

Understanding other people's minds is critical for understanding their behavior. Unfortunately for would-be behavior-understanders, people do not enjoy direct access to one-another's minds. Certain undeniable barriers preclude direct mind-to-mind connection—namely, skulls. Skulls are biologically useful, and it is not typically in people's best interest to breach them to better understand other's minds. Most people are aware of this, and so they instead employ a variety of mental state inference tools that psychologists have spent several decades researching.

This research has provided important information about how people infer others' transient thoughts, beliefs, desires, and emotions (for reviews, see Epley & Waytz, 2010; Fiske & Taylor, 2013; Mitchell, 2009). However, research into these different types of mental inferences has generally been subject to at least two limitations. First, these inferences are almost always studied one at a time. In real life, however, people infer mental states simultaneously (see Malle & Holbrook, 2012). In the present study, participants were required to mentalize simultaneously about people's emotions, beliefs, and intentions.

A second limitation is that existing research on the inference of transient mental states has largely ignored how motivations affect these inferences. This is perhaps surprising, since the importance of perceivers' motivations is widely appreciated and actively studied in most other areas of social cognition research, for example, stereotyping (Dijksterhuis, van Knippenberg, Kruglanski & Schaper, 1996; Fiske & Neuberg, 1990), prejudice (Roets, Van Hiel & Dhont, 2012), self-perception (Dunning, 1999), psychological essentialism (Keller 2005), and social cognitive neuroscience (Ames & Fiske, 2013b; Hughes & Beer, 2012) (see also Kruglanski, 1996; Kunda, 1990; Showers & Cantor, 1985; Stevens & Fiske, 1995).

Because motivations affect many social cognitive processes (as well as nonsocial judgments, e.g., Balcetis & Dunning, 2006), motivations could plausibly affect inferences of transient mental states as well. Some motivations may even be strong enough to affect basic mental state inference processes that people are typically assumed have mastered during childhood. The present research investigates this idea, and asks whether adults may make “childish” mentalizing errors under motivated conditions (Birch & Bloom, 2007; Diamond & Kirkham, 2005).

Because people experience many different motivations, motivated mentalizing likely arises from many causes. Here, however, we focus our investigation on the motivations that arise in response to intentional and unintentional harmful acts. Intent and harm are two concepts that are of particular importance to studies of motivated mentalizing precisely because of their importance to social psychological theory as a whole: Whether or not an action is intentional is arguably the first and most fundamental judgment that people make about others’ behavior (Malle, 2011; Malle & Holbrook, 2012). Whether those intentions are harmful or benevolent is arguably the first and most fundamental judgment that people make about other people (Fiske, Cuddy & Glick, 2007; Fiske, Cuddy, Glick, & Xu, 2002). It is unsurprising therefore, that, people are hypersensitive to harmful intents (Gollwitzer & Rothmund, 2009). This hypersensitivity may lead to currently-unappreciated biases in mental state inference processes, with implications for mentalizing accuracy.

Overview

We begin this investigation by reviewing three mentalizing tools that people typically master within the first several years of life: *false belief reasoning*, *second-order mental state reasoning*, and *mental state bootstrapping*. This review suggests these three tools—early-

emerging skills in human ontogeny—typically present no challenge (even in combination) for adults. This research provides a straightforward and testable prediction: that when people use these basic mental inference tools to impute affective states, they will be (appropriately) insensitive to the ancillary presence of bad intentions (the *intent-insensitivity hypothesis*). We then present an alternative hypothesis based in research suggesting that people are extremely sensitive to others' intents (the *intent-oversensitivity hypothesis*), and that when people detect harmful intent, they tend to have a moral response that involves the motivation to blame, condemn, and punish the harmdoer. To these observations, we add the recent finding that people who experience these motivations will sometimes perceive more harm than is actually present. Together, these observations give rise to a set of predictions from the intent-oversensitivity hypothesis that compete with those of the intent-insensitivity hypothesis. We then attempt to arbitrate between these two hypotheses.

False Belief Inference

Among mental state inferences, belief inferences are almost certainly the most-studied (Miller, 2009). As a matter of academic practice, researchers tend to focus on false beliefs, rather than true beliefs. This is because true beliefs are identical to reality. Therefore, in the case of true beliefs, it is impossible for social cognition researchers to know, experimentally, whether a particular social cognitive inference is due to the target individual's belief, or to the participant's own beliefs about reality (Dennett, 1978). False belief reasoning, by contrast, is thought to provide a relatively clean test of whether a person understands that human beings live in a world of thoughts, ideas, and beliefs, as well as actual world states.

The dominant paradigm for assessing false belief reasoning is variously referred to as the displacement task, the Sally-Ann(e) task, or simply the false belief task (Baron-Cohen, Leslie, &

Frith, 1985; Wimmer & Perner, 1983). The standard version of the task takes the following form: The participant (usually a 3-5 year-old) encounters two characters named Sally and Anne. Sally places a chocolate (or other object) in a basket, and then leaves the room. While Sally is away, Anne moves the chocolate from Sally's basket to a drawer. Participants answer four comprehension-check questions ("Which one is Sally?" "Which one is Anne?" "Where was the object when Sally left?" and "Where is it now?"), and finally, one critical question: "Where does Sally *think* the chocolate is now?" (or alternatively, "Where will she look for it when she returns?").

The correct response to this final question is that she will look where she left it (in the basket) rather than where it actually is (in the drawer). This deduction requires understanding that Sally is a mind-having entity, and that minds sometimes misrepresent reality. In this sense, passing the Sally-Anne task requires a conceptual distinction between mind and world (Dennett, 1978). Success in the Sally-Anne task has come to be taken as a marker for understanding the mind/world distinction more generally, and whether a child "passes" this task is often used as an important outcome measure in studies of children's social development (Dunn, Brown, Slomkowski, Tesla & Youngblade, 2001; Hughes & Dunn, 1998; Perner, Ruffman & Leekam, 1994; Youngblade & Dunn, 1995).

This empirical approach—using the Sally-Anne task to draw general inferences about whether a person possesses a theory of mind—has sometimes been criticized as being too narrow. These criticisms stem in part from the observation that the task focuses exclusively on beliefs, while ignoring other mental states, such as intentions, desires, and feelings (e.g., Bloom & German, 2000). One reason offered as to why researchers disproportionately study beliefs, rather than other mental states, is that the ability to understand beliefs is fundamental to the ability to

understand other mental states (Wellman et al., 2001). Knowing that Romeo *believes* that Juliet is dead gives us, “for free,” the inference that he *feels* sad. Notably (particularly for purposes of the present research), this emotional inference is equally valid regardless of whether Romeo’s belief happens to be true or false—that is, whether Juliet’s corpse is in the Capulets’ tomb, or Romeo’s mind.

Bootstrapping from Beliefs to Feelings

This inference process—by which people mentally bootstrap from knowledge of beliefs (Romeo believes that Juliet is dead) to inferences of emotions (Romeo is sad)—is an important part of how people normally predict and understand one another’s feelings (Baron Cohen, 1991; de Vega, Diaz & Leon, 1997). This point is highlighted in patient research on autistic populations. Autistic individuals are known to be profoundly impaired at understanding others’ beliefs (Baron-Cohen, et al, 1985; 1986). One study (Baron-Cohen, 1991) asked whether autistic individuals understand the causes of emotions. Three common causes of human emotion were examined: beliefs, desires, and situations. At test, people with autism were no different from matched controls at understanding emotions that were caused by situations or desires; however, they were significantly impaired in understanding others’ emotions when those emotions were caused by people’s beliefs. This research demonstrates that people with autism have difficulty understanding Romeo’s feelings—that is, they do not readily grasp that knowing how people construe a situation (what they *believe*) provides valuable insight into how they will *feel*.

Non-autistic individuals, on the other hand, intuitively grasp this belief→emotion connection. Children begin using beliefs to infer emotions at age 4-6 (Harris, Johnson, Hutton, Andrews & Cooke, 1989), almost as soon as they become capable of passing the false belief task

(age 4-5;⁴ Wellman, Cross & Watson, 2001). Humans without social cognitive deficits make use of other bootstrapping inferences as well (e.g., knowing someone's beliefs can help one guess at their goals; knowing someone's goals can provide clues as to what they feel, depending on whether or not those goals have been fulfilled). The present research focuses on the belief→emotion inference; however, studying the other bootstrapping inferences that people make in their everyday lives could be equally informative. In any event, with false belief reasoning and mental state bootstrapping in place, children are positioned to be able to think about others' minds (and to predict their behavior) in fairly sophisticated ways.

However a complication arises in that people whose mental states are being thought about are *themselves* often thinking about mental states. Therefore, mature social reasoning requires at least one more ability: the ability to understand that another person's mind can represent, not only facts, situations, and events, but also the contents of still *other* minds (the friar *supposes* that Romeo *believes* that Juliet is alive).

Second-Order Mental Inference

This kind of mental reasoning is called second-order mental-state reasoning, a capacity that is typically in place by age 5-9 (Miller, 2009; Perner & Wimmer, 1985; Sullivan, Zaitchik & Tager-Flusberg, 1994). Its reported developmental trajectory varies partly as a function of assessment method (and its precise point of emergence is not particularly relevant to the present

⁴ In recent years, several challenges to this traditional view have emerged, suggesting that younger children, or even infants, may possess rudimentary false belief understanding (see Baillargeon, Scott, & He, 2010; Sodian, 2011, for reviews and discussion). However, the question of when children begin to understand beliefs is not particularly relevant to the present investigation.

investigation). The most frequently-used methods are structurally similar to the Sally-Anne (first-order false belief) task described above. Thus, to return to Sally, Anne, and the perpetually moving chocolate, “Anne *intends* to hide Sally’s chocolate” constitutes a first-order mental state inference. “Sally *supposes* that Anne *intends* to hide Sally’s chocolate” constitutes a second-order mental state inference.⁵ Second-order mental state reasoning underlies much of people’s everyday social cognition, since it is difficult to make sense of other people and their behavior without understanding that people think about others’ thoughts, beliefs, feelings, and desires.

While this thinking-about-minds-thinking-about-minds process may sound complex, second-order mental state reasoning poses no challenge to adult humans (Dunbar, 1998). It has been theorized that the human brain evolved under significant pressure to become highly efficient at performing these computations; and in fact, most adults can track second, third, or even higher-order mental states without much difficulty (the friar *believes* that Romeo *knows* that Juliet *intends* to fake her own death; Kinderman et al., 1998; Stiller & Dunbar, 2007).

Synthesis: Inferring Emotions via Second-Order Belief Reasoning

The theater audience understands why Romeo is sad, even though the audience knows that Romeo has a false belief about Juliet’s “death.” If you were able to make sense of the previous sentence, then you have successfully integrated the three mental inference tools reviewed above (false belief reasoning, bootstrapping from beliefs to emotions, and second-order mental state reasoning). You likely accomplished this synthesis with very little effort, and would be able to combine these elements in different ways to make other inferences. For example, suppose that Romeo believed that Juliet was only pretending to be dead, in order to get Romeo to

⁵ Different researchers employ different counting systems to describe mental inference levels.

Here, we follow the system first introduced by Perner and Wimmer (1985).

leave her alone forever. Regardless of whether or not Romeo's beliefs about Juliet's intentions happen to be correct, you might well infer that Romeo would be very upset.

This apparently straightforward mental inference is structurally identical to the inference that participants were asked to make in the following three experiments, and highlights a long-understood principle of human psychology: when people experience harm, their emotions depend, in part, on whether or not they believe that the harm was *intended* (Holmes, 1881; Vangelisti & Young, 2000). Put another way, it feels bad to be harmed, but worse to be *victimized*—that is, to be harmed intentionally. If we know that someone believes that they were harmed intentionally rather than unintentionally, we should have a social-cognitive “leg up” in predicting how they will feel.

To relate this observation to the classic Sally-Anne paradigm, suppose that one wanted to predict Sally's feelings. Most people would infer (probably correctly) that Sally would be fairly upset if she *believed* that her friend Anne *intentionally* hid the chocolate out of pure spite. However, in an identical situation in which Sally *believed* that her friend Anne had *accidentally* made the chocolate hard to find by moving it to a safer location, most people would infer (again, probably correctly) that Sally would be less upset than in the malicious intent case. You may have noticed that *Anne's actual intentions are completely irrelevant to Sally's feelings*. If Sally *falsely* believes that Anne meant to hurt her, she will feel terrible. If Sally *correctly* believes that Anne meant to hurt her, she will feel equally terrible (assuming that Anne's actual behaviors are the same).

Sensible as this analysis may be, actual social perceivers may not always make this mind/world distinction so cleanly. That is, socially attuned minds may find it difficult to discount Anne's true (and completely irrelevant) intentions, despite sophisticated cognitive metalizing

systems that apparently make such discounting relatively easy in the absence of motivation. This may represent a particular instance of a broader class of effects: a “curse of knowledge” effect (Camerer, Loewenstein, & Weber, 1989).

Competing Hypotheses

The review just presented gives rise to two competing hypotheses: the intent-insensitivity hypothesis and the intent-oversensitivity hypothesis. The intent-insensitivity hypothesis is rooted in decades of research showing that (a) false-belief reasoning, (b) belief→emotion mental state bootstrapping, and (c) second-order mental state reasoning are facile, necessary, and early-emerging components of everyday social cognition (e.g., Cosmides & Tooby, 2013; Kinderman et al., 1998; Miller, 2009; Perner & Wimmer, 1985; Stiller & Dunbar, 2007; Sullivan et al., 1994). This well-established perspective offers a clear prediction with respect to the modified version of the Sally-Anne task described above. On this account, adults can separate their own beliefs from Sally’s false beliefs (false belief reasoning) about Anne’s intentions (second-order mental state reasoning), and can then use those beliefs to make inferences about Sally’s emotions (mental state bootstrapping). Participants’ reasoning should therefore align with the mind/world distinction described above: Sally’s feelings depend on her own beliefs about Anne’s intentions (helpful vs. harmful), but not what Anne’s true intentions happen to be. Put another way, for purposes of inferring Sally’s emotions, it is appropriate to be insensitive to Anne’s true intentions.

An alternative hypothesis, which we will call the intent-oversensitivity hypothesis, states that people will be unable to fully discount Anne’s bad intentions (despite “knowing” that they should do so). Under this hypothesis, participants’ own knowledge of Anne’s true intentions will (inappropriately) influence their judgments of Sally’s affect. Specifically, they will report Sally

as being more hurt when Anne's actual intentions happen to be malicious rather than benign, even if the participants know that Sally (the victim) has no idea (or the wrong idea) about whether Anne's intentions were malicious or benign. This hypothesis draws indirect support from the observation that, across a wide range of theoretical perspectives, data have emerged showing that people are extremely sensitive to the presence of intentional harms (Alicke, 2000; Darley & Pittman, 2003; Gebotys & Dasgupta, 1987; Gollwitzer & Rothmond, 2009; Gray & Wegner, 2010; Haidt & Kesebir, 2010; Knobe et al., 2012; Shaver, 1970).

Further support comes from repeated demonstrations that people do often fail to discount their own privileged knowledge when attempting to predict others' thoughts, feelings and behavior (Birch, 2005; Birch & Bloom, 2007; Camerer, et al., 1989; Epley, Keysar, Van Boven, & Gilovich, 2004). For example, participants in one frequently cited experiment (Newton, 1990) tapped out the rhythm of a famous song (e.g., "Happy Birthday") with their finger while their partners attempted to guess the song. The people tapping out the song predicted that their partners would be able to guess the song correctly 50% of the time. In reality, the song was guessed correctly about 3% of the time. The "tappers" reported being able to clearly hear the music and lyrics in their heads as they tapped out the song; but in estimating the contents of their partners' minds, the "tappers" failed to appropriately discount these experiences, to which their partners did not have access.

Yet people do not always show this kind of egocentrism in reporting the likely contents of other minds—again, in the standard false belief task adults successfully discount their own privileged knowledge of the hidden object's location nearly 100% of the time. Whether and to what extent this kind of egocentric metalizing emerges is often conceptualized as a product of (a)

the perceiver's motivations and (b) their available cognitive resources (Epley, et al., 2004).⁶ A more general version of this principle may also be applied: People arrive at different judgments (about the minds of people, the value of toasters, the prestige of schools, and everything else) as a function of what information they use and what decision rules they apply; and different information and rules are brought to bear depending on the motivations and available cognitive resources of the perceiver (Kunda, 1990). Generally speaking, a lack of cognitive resources (e.g., low intelligence, high task demands) will bias processing toward simpler computations (often based on heuristics; Lin et al., 2010). Motivations induce more specific directional biases, causing people to selectively search memory for information, heuristics and decision rules that will allow them to reach the desired conclusion (Kunda, 1990). This selection processes usually occurs outside of awareness, such that people are unaware that their selection of information and processing strategies may be biased (the "illusion of objectivity" discussed in Chapter 1; Pyszczynski & Greenberg, 1987). One result of this is that, even if a piece of information is logically irrelevant (Anne's privately-held intention) to the judgment at hand (Anne's feelings), the irrelevant information may still give rise to motivations (blame motivation) that subtly affect people's judgments.

Following this reasoning, we predict that oversensitivity to Anne's harmful intentions will lead to inflated perceptions of Sally's emotional pain. While this view appears to be in tension with an established literature on false beliefs, belief-to-emotion mental state bootstrapping, and higher-order mental state reasoning, the research reviewed in Chapter 1

⁶ Task complexity, perceiver-target similarity, and the perceiver's mood play a roll as well (Ames, 2004; Ames, Jenkins, Banaji, & Mitchell, 2008; Birch & Bloom, 2007; Converse, Shuhong, Keysar & Epley, 2008; Lin, Keysar & Epley, 2010), though at least some of these factors could be conceptualized as exerting their influence indirectly by affecting motivation and or resource availability.

provides one additional piece of supporting evidence: Participants consistently saw intentional (non-emotional) harms as worse than identical unintentional harms—even when the harms were objectively quantifiable, and this effect appeared to be mediated by blame motivation. The hypothesis that these same motivations forces could influence people’s affective mentalizing has not yet been tested. This is the primary hypothesis we aim to test here.

It is unclear whether perceptions of affect would be influenced in the same way as perceptions of other kinds of harms, given that the cognitive and neural mechanisms used to think about beliefs and emotions are substantially different from those used to calculate damages in dollars and cents (Amodio & Frith, 2006; Cosmides & Tooby, 2013; Hooker, Verosky, Germine, Knight & D’Esposito, 2008; Mitchell, 2009; Saxe & Kanwisher, 2003). In addition, there has been virtually no research concerning how perceivers’ motivations impact the perception of others’ affect. Moreover, it is thought that the human brain developed under relentless evolutionary pressure to successfully track who knows what about whom (Cosmides & Tooby, 2013), which would prevent the inappropriate application of privileged knowledge. While this successful separation may be disrupted by working memory load and/or highly complex task designs (e.g., Lin et al., 2010, Birch & Bloom, 2007), we do not impose either of these conditions here. Rather, we simply suggest that, for all of its sophistication, the human mentalizing system is not immune to motivation (Epley et al., 2004), and that motivated mentalizing might therefore lead to empathic errors.

Participants

Participants were recruited using Amazon Mechanical Turk, and then redirected to an experimental website where they completed the study. Participants were paid between \$0.30 and \$0.65 for their participation in these studies. Recruitment was limited to the United States. As

noted in Chapter 1, current research suggests that samples recruited via Mechanical Turk are more demographically and cognitively representative of national distributions than are traditional student samples, and that data reliability is at least as good as that obtained via traditional sampling (e.g., Berinsky, Huber & Lenz, 2012; Buhrmester, Kwang & Gosling, 2011; Horton, Rand & Zeckhauser, 2011; Mason & Suri, 2012; Paolacci, Chandler & Ipeirotis, 2010). Recruitment was limited to persons living in the U.S., and who had at least a 95% approval rating on MTurk. In addition, participants who had completed any previous study in this line of work were automatically prevented from accessing the experimental website via code developed by Pe'er, Paolacci, Chandler and Mueller (2012).

Study 1

Method

Because the Sally-Anne false belief task is the dominant paradigm for assessing whether people can track the contents of other minds, we began our investigation with a modified version of this paradigm. We retained the structure of the task, while expanding the paradigm to provide a test of empathic mentalizing (similar to Hooker et al., 2008), and making adjustments for an adult population (e.g., the characters were adults and the object in question was a ring instead of a chocolate or toy; see Appendix D for complete text). The scenario described Sally and Anne washing dishes together. Sally took off her ring and put it in the soap dish. She then left to talk on the phone. While Sally was gone, Anne moved the ring to a drawer, such that Sally would not be able to find it, which would probably make Sally upset. In the *intentional harm* condition Sally intended to make Sally upset. In the *unintentional harm* condition, Sally did not intend this (she only intended to move the ring to a safer location).

The classic false belief task almost always includes pictures or puppets to ensure narrative comprehension (Wellman et al., 2001). While this may not be necessary for adult populations, we nonetheless included illustrations depicting the scene as described in the vignette, making it clear exactly when Sally left the room (and thus, when she ceased to have updated knowledge of the hidden object's location). The illustrations were identical across the two conditions (see Appendix E for complete set). After participants had read the narrative and viewed the corresponding illustrations, they answered the four standard comprehension questions ("Which one is Sally?" "Which one is Anne?" "Where was the ring when Sally left the room?" and "Where is the ring now (where the story left off)?"), and finally, the standard false belief question, "Where does Sally think that the ring is now (where the story left off)?"

Consistent with existing research, we anticipated that adult participants would be near ceiling on passing the false belief performance (i.e., they would almost always answer correctly when asked where Sally thought the ring was). This would traditionally denote a general capacity to reason about others' minds independently from one's own privileged knowledge. However, consistent with the *intent oversensitivity hypothesis*, we expected a novel dissociation between false belief reasoning and empathic mentalizing: participants would pass the false belief task, but fail the affective inference task. Specifically, we predicted that participants would judge Sally to be more emotionally hurt when participants knew that Anne tried to hurt her intentionally, even though Sally was blind to Anne's intentions. To assess participants' judgments of emotional harm, we asked: "How bad (emotionally) do you think Sally will feel when she sees that the ring is missing?" Participants gave their response on a 0-100 slider scale, anchored with "Not at all" and "The worst that she could feel (emotionally)."

Next, we assessed the proposed mediator: blame motivation. As in Chapter 1, we assessed this construct via a three-item index ($\alpha = .86$): “To what extent would you want Anne to be punished for moving Sally’s ring?” “To what extent do you think Anne deserves blame for moving Sally’s ring?” “To what extent do you think Anne deserves moral condemnation for moving Sally’s ring?”

Results

Reading checks. Psychometric investigations show that the reliability and statistical power of a dataset can be significantly improved by employing simple reading checks to eliminate a minority of participants who are merely “clicking through” (Oppenheimer, Meyvis & Davidenko, 2009). Participants were therefore excluded from analysis if they failed one or more of the standard comprehension checks (“Which one is Sally?” “Which one is Ann?” “Where was the object when Sally left?” “Where is it now?”). Thirty-four such participants were excluded—consistent with previous research suggesting that a subpopulation of “participants” are not actually participating in psychological studies, and that this is usually more common in Mechanical Turk samples than in university samples obtained in the lab. This same research shows, however, that this population can be usually be identified and removed using basic reading/comprehension checks, increasing power and reliability, and yielding Mechanical Turk samples that are identical to or only slightly different from laboratory samples in terms of task performance (Horton et al., 2011; Mason & Suri, 2012; Oppenheimer et al., 2009; Paolacci et al., 2010). Based on the same principle, one additional participant was excluded for apparently reading the scenario in less than 5 seconds (following the recommendations of Mason & Suri, 2011).

Belief inference. Of the remaining 116 participants, only two failed the false belief task

(reporting that Sally thought that the ring was in the drawer to which Anne had moved it, rather than by the sink where Sally herself had left it). The inclusion or exclusion of these two participants made no difference in the results (the analysis reported below includes them).

Intent. Perceived intentionality was manipulated successfully, with participants viewing the harm as more intentional in the intentional condition ($N = 54$, $M = 87.3$, $SD = 24.0$) than in the unintentional condition ($N = 62$, $M = 2.9$, $SD = 10.1$, $t(69.2)=24.1$, $p < .001$, $d=5.03$, corrected for unequal variances). Participants answered the following question: “To what extent did Anne intend to hurt Sally emotionally by moving the ring?” (0 = *Did not intend this at all*; 100 = *Completely intended this*) Ames & Fiske, 2013a).

Harm. Our key prediction was that participants would judge Sally’s feelings to be more hurt in the intentional harm condition than in the unintentional harm condition. This prediction was confirmed, $M_{(intentional)} = 43.4$, $SD = 22.9$; $M_{(unintentional)} = 34.2$ $SD = 22.3$, $t(114)=2.19$, $p = .03$, $d=0.41$. This result violates the straightforward logic of the intent insensitivity hypothesis, under which perceivers who correctly track Sally’s beliefs about the ring should also correctly track Sally’s beliefs about Anne’s intents, and use those beliefs (but not the intents themselves), to infer Sally’s feelings.

Blame motivation. While Anne’s intents are not *structurally* relevant to participants’ emotion inferences, past research suggests that they might be *motivationally* relevant, and that blame motivation might explain why the intent to do harm could inflate the amount of harm perceived. To test whether this motivational account could potentially describe the observed data, we examined blame motivation as a potential mediator. First, as predicted, participants were more motivated to blame Anne when she had upset Sally intentionally, $M = 72.7$, $SD = 21.1$, rather than unintentionally, $M = 12.5$, $SD = 13.5$, $t(114) = 18.6$ $p < .001$, $d = 3.40$ (each item of

the subscale was also significant when tested individually, $ps < .001$). Moreover, blame motivation mediated the effect of perceived intentionality on magnitude of perceived harm, 1,000 bootstrapped samples, 95% CI [.176, .889], completely standardized, [.090, .457] unstandardized, Sobel: $Z = 2.88$, $p = .004$. The direct effect of intent on perceived harm dropped to nonsignificance ($p = .10$) after accounting for the mediator (i.e., “full mediation”; but see Rucker, Preacher, Tormala & Petty, 2011, for a principled critique of the conventional distinction between “full” and “partial” mediation).

As in Chapter 1, bootstrapping mediation and Sobel tests strongly supported the predicted model (intent→blame motivation→perceived harm) over the alternative (and arguably more intuitive) model (intent→perceived harm→blame motivation; see Appendix F for full report of mediation analyses). This finding parallels similar analyses examining mediation for other kinds of harm (Chapter 1/Ames & Fiske, 2013b). It also comports well with observations by Darley and colleagues that, in the case of intentional harms, moralistic, *retributive motivations*—rather than the actual *amount of harm*—primarily determine the amount of punishment that observers deem appropriate (Carlsmith, Darley & Robinson, 2002; Darley & Gromet, 2010; Darley & Pittman, 2003).

Discussion

Participants passed the false belief task—yet, at least in some sense, they failed the emotional inference task. They passed the false belief task in the sense that they successfully separated Sally’s beliefs from their own privileged knowledge (which Sally did not possess). They failed the emotional inference task in the sense that their inferences of Sally’s emotions were influenced by their own privileged knowledge (which Sally did not possess). This did not have to be the case—there are many instances in which people successfully inhibit their own

knowledge, and many in which they do not (see Introduction). That they should fail to do so here introduces an uncomfortable tension with the literature reviewed in the Introduction demonstrating that adults can easily separate their own knowledge from others' beliefs, and base their mental inferences on the latter, rather than the former, when appropriate. One possible explanation for this apparent divergence is that this literature tends not to consider how perceivers' motivations might influence their inferences of transient mental states. Motivation to blame could have caused participants to have distorted views concerning the harm that was done (in this case, emotional harm). Mediation analyses are consistent with this explanation for the present data; however, such analyses are hardly definitive, and alternative explanations also exist.

One such alternative arises from the fact that (following the traditional Sally-Anne task format) the vignette did not explicitly define Sally's beliefs about Anne's intentions. Therefore, participants could plausibly have *inferred* different beliefs across the two conditions. For example, since participants in the intentional harm condition read about one person hurting the other intentionally, they might have inferred a more strained relationship between Sally and Anne overall. If Sally's relationship with Anne was more strained, one might reason that Sally would suspect negative intent on Anne's part—while this inference would be less likely in the unintentional harm condition. Under this account, the results of Study 1 do not constitute evidence for the intent-oversensitivity hypothesis, since the intent-oversensitivity hypothesis predicts a difference in empathic inferences *despite* an identical explicit understanding of the victim's beliefs.

In Study 2, we aimed to address this alternative account. Specifically, we told participants exactly what belief the victim held about the harm (that it was intended). Handing participants this information would seem to give them an advantage in determining how the victim would

feel, since, as discussed earlier, people know that people usually feel worse when they believe they've been harmed intentionally vs. unintentionally. If the alternative account presented above (different inferences about the victim's beliefs) explains the results of Study 1, providing this belief information explicitly should eliminate the effect of intent on perceived harm. However, if the intent-oversensitivity hypothesis correctly characterizes the psychological processes underlying the observed effect, making the belief information explicit should make no difference.

Another possible objection to Study 1 is that the scenario described represents a special case, insofar as it describes describe a violation of trust between two close individuals. Such trust violations differ from other kinds of emotional harms in important ways, producing qualitatively distinctive negative psychological reactions (see Finkel, Rusbult, Kumashiro & Hannon, 2002, for review). Therefore, the results of Study 1 may not generalize beyond trust violations between close parties. However, if these results are due to blame motivation more generally, rather than to psychological factors endemic to trust violations, the same results would be expected for emotional harms that are not trust violations. Study 2 therefore employs a vignette in which no trust violation takes place.

A third possible objection to Study 1 is that the Sally-Anne task (as used here and elsewhere) presents an unnaturally sterile and constrained mental inference situation (for discussion, see Bloom & German, 2000). This simplicity is by design; however it exemplifies a general weakness of many vignettes used in psychological research: Brief, structured scenarios provide good experimental control, but may cause perceivers to treat them like games or puzzles, rather than as simulated social situations. Unlike more developed narratives, which elicit genuine emotions and social impulses (de Vega & Leon, 1997) vignettes like the Sally-Anne scenario may fail to evoke the very psychological responses ostensibly under investigation. Still, people

tend to understand the social world in narrative terms (Murray, 2007), and written narratives provide a useful tool for sensitive areas of inquiry (such as intentional harmdoing) that might not be assessable via other methods.

In Studies 2 and 3, we sought to mitigate some of the weaknesses of the vignette approach while retaining its strengths. To this end, we generated lengthier and more engaging narratives for Studies 2 (short story excerpt) and 3 (video). These more expansive narratives also provide a more conservative test of our hypothesis in at least one sense: they provide several valid sources of information upon which participants may base their responses (e.g., the social setting, descriptions (Study 2) or pictures (Study 3) of the characters). If changing one small component of the stimulus materials (in this case, one character's intentions) produces an effect despite the presence of other valid sources of information, it seems likely that the element that was changed was causally important.

Study 2

Method

Participants read an extended narrative about Cole, a college student, and his love interest, Veronica. The narrative was presented as an excerpt of a larger piece of fiction. Participants were randomly assigned to either the intentional condition or the unintentional condition, which specified whether Veronica hurt Cole's feelings intentionally or unintentionally. In the unintentional condition, the story excerpt read as follows:

Cole had barely stopped thinking about her since the party. The way her hair looked, what her voice sounded like, how she fidgeted with the penguin-shaped charm on her

bracelet... Then, as he was coming out of his dorm that morning, he spotted her sitting with a friend.

Veronica. Even her name intimidated him. It gave him goosebumps when he dared whisper it to himself. He sometimes spoke it under his breath when he was sure no one would hear. Like a prayer—one that invoked the private thrill of a thousand tiny pinpricks on his skin. He had kept imagining (kept longing for) unlikely situations that would give him an excuse to talk to her again. Of course, when he actually saw her sitting outside of his dorm that morning, his first instinct was to immediately run back inside and hide under his bed until she left. But then he saw her look in his direction, and he jerked his hand up in a hopeful wave. He smiled at her, his heart in his throat.

Veronica did not see him wave, and didn't wave back. She was a pretty sophomore (though, it must be said, not as pretty as Cole imagined her to be). She danced ballet, did very well in her classes, and got a fair amount of attention from boys. She wasn't especially interested in Cole, but she never intended to hurt him by not responding to his wave. She turned toward her friend, so that her back was facing Cole, and she started talking about a party she was going to that night. Cole felt pretty sure that Veronica had seen him and was ignoring him on purpose. His legs seemed to stop functioning properly for a moment, and he just stood there blinking. Then he seemed to regain the ability to walk, and made his way to class--though he didn't absorb much of the lecture.

The intentional harm condition was identical except for a few key words in the third paragraph:

*Veronica **saw** him wave, **but** didn't wave back. She was a pretty sophomore (though, it must be said, not as pretty as Cole imagined her to be). She danced ballet, did very well in her classes, and got a fair amount of attention from boys. She wasn't especially interested in Cole, **and she intentionally didn't respond to his wave so that he would be hurt, and leave her alone forever.** She turned toward her friend, so that her back was facing Cole, and started talking about a party she was going to that night.*

A key feature of this experiment is that Cole's beliefs about Veronica's intentions were stated explicitly: In both conditions, he believed that Veronica was snubbing him. This construction rules out the first alternative interpretation presented for Study 1 (different belief inferences across conditions), and should have caused participants to converge on the same harm ratings. After all, Cole's feelings really would arise from his beliefs, regardless of whether those beliefs happened to be true. Nonetheless, we predicted that people's extreme sensitivity to harmful intentions would render them unable to fully discount Veronica's true intentions—even when Veronica's true intentions were, in fact, irrelevant.

Results

Reading comprehension check. Participants responded to a single reading comprehension check question: “Did Veronica see Cole wave at her?” (yes/no). This question was presented immediately after the vignette; 19 participants failed to answer correctly, and one additional participant apparently read the story in less than five seconds, suggesting that a non-

negligible minority failed to read the story carefully. However, the exclusion or inclusion of these 20 participants did not impact the results (see footnote⁷ for full comparison).

Intent. Participants saw the emotional harm as more intentional in the intentional condition, $N = 81$, $M = 75.3$, $SD = 29.9$, than in the unintentional condition, $N = 78$, $M = 7.5$, $SD = 17.2$, $t(128.85)=17.61$, $p < .001$, $d=2.78$, corrected for unequal variances. This was assessed by the question: “To what extent did Veronica *intend* to hurt Cole's feelings?” (0 = *Did not intend this at all*, 100 = *Completely intended this*).

Harm. Despite the fact that all participants were explicitly told how Cole construed the situation (as intentional harm), participants still judged Cole's feelings as significantly more hurt when the participants themselves knew that the harm was intentional ($M = 74.7$, $SD = 20.6$) rather than unintentional ($M = 65.7$, $SD = 23.7$), $t(157)=2.58$, $p = .01$ $d=.41$.

Replication. The wording of Cole's beliefs, “Cole felt pretty sure that Veronica had seen him,” is potentially problematic for this study. Specifically, it is possible that participants in the *intentional harm* condition read “pretty sure” to mean something like 95% confident, while participants in the *unintentional harm* condition read “pretty sure” as something like “65% confident.” We therefore re-ran the study replacing the words “Cole felt pretty sure” with “Cole was sure.” Results were identical to those observed with the original wording, $M_{(\text{intentional})} = 69.4$, $M_{(\text{unintentional})} = 60.1$, $t(181) = 2.96$, $p = .003$, $d = .44$. In addition, we asked: “True or false:

⁷ Paralleling the inclusive analysis described in the main text, an exclusive analysis found that participants saw more harm in the intentional condition ($M = 75.5$, $SD = 19.9$) than in the unintentional condition ($M = 66.6$, $SD = 23.4$, $t(137)=2.4$, $p = .02$ $d = .41$). Results for intent were likewise statistically equivalent: intentional: $M = 76.7$, $SD = 28.8$; unintentional: $M = 3.6$, $SD = 9.2$, $t(97.95)=21.18$, $p < .001$, $d = 3.42$, corrected for unequal variances.

According to the story, Cole believed that Veronica was ignoring him on purpose,” and excluded participants who answered incorrectly. Again, this did not change the results $M_{(\text{intentional})} = 69.4$, $M_{(\text{unintentional})} = 61.6$, $t(170) = , p = .01$, $d = .38$. It therefore appears that the two conditions were identical from Cole’s perspective, yet participants, with their privileged knowledge of Veronica’s intentions, still judged that Cole would have a different emotional response.

Discussion

The fact that participants made different emotional inferences despite identical knowledge of Cole’s beliefs provides a more rigorous test of our hypothesis. On the intent-insensitivity account, analysis of Cole’s emotional state should be relatively straightforward: There were no differences in Cole’s subjective reality across conditions—therefore his emotions should be the same. However, this straightforward line of reasoning did not comport with people’s actual judgments.

The intent oversensitivity hypothesis predicts the pattern of results that was actually observed; however, as with Study 1, there are still possible objections to this account of the data. Perhaps the most important is that, although Veronica’s actions were *described* identically across the two vignettes, participants may still have *imagined* her behaving differently across the two conditions. People tend to visualize narratives as they read them, and participants in Study 2 may have invented affectively relevant details that differed between the two scenarios. By design, the narrative employed in this experiment provided scant opportunity to invent such relevant differences in the Cole-Veronica interaction between conditions (she stayed turned away from him, making differences in her facial expression irrelevant; and the complete absence of dialogue eliminates paraverbal considerations). Still, it is possible that the participants in the intentional harm condition consistently invented more insulting ways for Veronica to turn her back to Cole

(imagining, for example, a condescending hair flip in the intentional harm condition, but not in the unintentional harm condition). This detail, in turn, could plausibly have led to Cole's feelings being more hurt.

While this is a possible explanation for the results of Study 2, it cannot explain the results of Study 1. How could a ring be more or less insulting in its absence? Still, as we have noted, Study 1 has its own limitations in that it failed to explicitly state the victim's belief about the intentionality of the harm. Study 3 attempts to address both of these issues simultaneously. First, it presents the vignette both visually and in words, so participants should be unlikely to visualize the story differently. Second, just as Study 2 explicitly stated that the victim thought the harm was intentional, Study 3 explicitly states that the victim thought that the harm was unintentional. In addition, because Study 2 lacks a measure of blame motivation, no mediation analysis could be conducted (though participants' spontaneous comments appear consistent with blame motivation interpretation⁸). Study 1 does provide mediational evidence, and this analyses closely resembled previous instances in which blame motivation inflated other kinds of harms (Chapter 1). Still, it seemed worthwhile to test whether this mediational account describes the observed data for emotional harms outside the limited scope of the Sally-Anne task. Study 3 provides this test.

Study 3

Method

In this study, the emotionally hurtful event was portrayed via a YouTube video [http://www.youtube.com/watch?v=t_r24zMMec]. The video was presented as an installation of a weekly comedy vlog (video web log) created for entertainment purposes. The animated

⁸ E.g., "I hate Veronica;" "Girl sounded like a jerk."

video (approximately three minutes in length) focused on a contemporary romantic plotline. This plotline revolved around a protagonist named Eric, who attended a party. Eric's co-worker also attended the party, and posted a picture of Eric with a girl (in fact, Eric's cousin) to Facebook—either accidentally or intentionally implying romantic involvement, and thus creating strife between Eric and his actual romantic partner, who was not at the party. Eric heard about the post from another individual, and had no reason to suspect ill intent on the part of his coworker.

The visual component of the video was identical across conditions. The audio component (narrator voiceover; 588 words in each condition) was nearly identical, except for a few phrases (totaling 32 words) indicating whether the harm was being inflicted intentionally or accidentally (e.g.: unintentional harm condition: “...*Needless to say, she drew her own conclusion from this post—which is to say, a totally different conclusion from what Doug intended;*” intentional harm condition: “...*Needless to say, she drew her own conclusion from this post—which is to say, exactly the conclusion that Doug intended...*”).

After viewing the video, which showed Eric's and the other characters' affective and behavioral responses to the events (completely identical across conditions, both visually and auditorily) participants were asked to judge how much emotional harm Eric's co-worker had caused Eric by posting the picture to Facebook. As in Study 1, participants also completed measures of the co-worker's intent to cause harm, and the three-item measure of blame motivation.

Results

Viewing check. Timing data were collected to ensure that participants watched the video and to measure how long participants spent on the page that played the video. The timing element was not triggered in six cases, suggesting either that these participants skipped the video

portion of the task, or that an unknown data recording error occurred. These six participants were not included in the analysis. All of the remaining 97 participants appeared to have watched the entire video, ($M = 207.3s$, $SD = 19.8s$).

Intent. Intent was manipulated successfully, with participants seeing the harm as more intentional in the intentional harm condition ($N = 49$, $M = 73.2$, $SD = 32.8$) than in the unintentional harm condition $N = 48$, $M = 4.3$, $SD = 11.0$, $t(58.77) = 14.0$, $p < .001$ $d = 2.83$. This was assessed by the question: “To what extent did Doug *intend* to hurt Eric?” (0 = *Did not intend this at all*, 100 = *completely intended this*).

Harm. As with the first two studies, the intentional and unintentional harm conditions compared the same act committed by the same individual resulting in the same harms to the same person. Moreover, the victim’s (Eric’s) subjective understanding of the harm was equivalent across cases. Nonetheless, participants judged the harm as worse when the harm was intended, despite the protagonist’s equivalent beliefs ($M_{intentional} = 71.5$, $SD = 21.1$, $M_{unintentional} = 62.6$, $SD = 20.1$, $t(95) = 2.07$, $p = .04$ $d = .42$). The magnitude of the effect was nearly identical to the magnitudes observed in the previous two studies (see general discussion), both in terms of d , and as observed on 0-100 scale.

Blame motivation. As predicted, participants were more motivated to blame Eric’s coworker when he had harmed Eric intentionally ($M = 70.0$, $SD = 24.1$) rather than unintentionally ($M = 16.7$ $SD = 17.8$). Each item of the scale ($\alpha = .87$) was also significant individually, $ps < .001$. Moreover, blame motivation again mediated the effect of perceived intentionality on magnitude of perceived harm, regardless of whether this was assessed via bootstrapped confidence intervals (1,000 bootstrapped samples, 95% CI [.190, .651], completely standardized, CI [.089, .350] unstandardized) or the Sobel test ($Z = 3.12$, $p = .002$). Also as with

Study 1, the effect of intent on perceived harm dropped to nonsignificance ($p = .30$) when the mediator was accounted for. Finally, the proposed model, intent→blame motivation→harm, was again strongly supported over the alternative intent→harm→blame motivation model (see Appendix F for full report of mediation analyses).

Replication. As with Study 2, we replicated the paradigm with an explicit statement about the victim's beliefs. The audio component of the narrative was identical to that of the first video; however the replication version showed Eric calling and confronting Doug, who apologized ("I'm so sorry man. It was an accident!"). To eliminate ambiguity about whether Eric believed that Doug was telling the truth about it being an accident, the words "believes him" appeared next to Eric while Doug spoke. Doug then said, "I'll take it [the post] down," to which Eric responded, "Okay."

None of this altered the results obtained from the original video. Eric was still judged to be more hurt in the intentional condition ($M = 70.5$) than in the unintentional condition ($M = 61.0$, $t(167.6) = 2.72$, $d = .40$, adjusted for unequal variances), and were more motivated to blame Doug in the intentional condition ($M = 67.9$) than in the unintentional condition ($M = 23.7$, $t(174.3) = 12.6$, $d = 1.88$, adjusted for unequal variances). The mediation results were also identical (1,000 bootstrapped samples, 95% CI [.156, .539], completely standardized, CI [.086, .322] unstandardized); Sobel $Z = 3.44$, $p = .0006$). Also as with previous studies that assessed statistical mediation (Studies 1 and 3 from Chapter 1, Study 1 of this chapter), the effect of intent on perceived harm dropped to nonsignificance ($p = .85$) when the mediator was accounted for. Finally, also as in the previous studies, the proposed model, intent→blame motivation→harm, was supported over the alternative intent→harm→blame motivation model (see Appendix F for full report of mediation analyses).

Discussion

Again, participants failed to mentalize in the straightforward manner predicted by the intent-insensitivity hypothesis (appropriately discounting privileged knowledge of the harmdoer's intent). Instead, privileged knowledge of the harmdoer's intent elicited motivation to blame, which in turn influenced the empathic inference process, such that people attributed greater emotional suffering to a victim of intentional (vs. unintentional) harming, despite the two harms actually being equivalent from the victim's perspective.

General Discussion

Across three experiments (plus two replications), victims were judged to feel worse when someone hurt their feelings intentionally rather than accidentally. This would be entirely sensible if the victims *knew* that they were being hurt intentionally. The deliberateness of an attack is itself painful, and even physical pain hurts more when it was administered intentionally (Gray & Wegner, 2008). Some researchers have explained the relationship between hurt feelings and hurt bodies by suggesting that the physical pain system was “co-opted” during our evolutionary history, so that signals of “hurt” could be applied to socially threatening situations, thus helping us to avoid them. These researchers point out that physical and social pain both rely on mu-opioid-related signaling, recruit a set of common brain areas, and can be alleviated by common interventions (acetaminophen) (e.g., Chiang, Eisenberger, Seeman, & Taylor, 2012; Eisenberger, 2012; Eisenberger, Lieberman & Williams, 2003). Regardless of the underlying phenomenology, it seems clear that people usually feel worse when their feelings are hurt on purpose rather than on accident.

However, the victims in these studies did not know whether they were being hurt on purpose or on accident. The fact that victims themselves could not make this distinction is underscored by the replications of Studies 2 and 3: participants continued to apply their own privileged knowledge of harmdoer intent, no matter how clearly it was made that victims did not share this knowledge.

If victims themselves could not make this distinction (between being harmed intentionally or unintentionally), then this distinction cannot have impacted the victims' feelings. Adult observers clearly understand this and mentalize accordingly when performing standard mentalizing tasks. However, when observers mentalize from a motivated state, the typically solid boundary that distinguishes mind from world may begin to weaken, and information that would otherwise be blocked by this boundary may begin to begin to leak through, and (inappropriately) influence mentalizing judgments in non-normative ways. This idea aligns well with past research showing that people can reflectively distinguish between one's own belief and another's belief, yet fail to reflexively apply that distinction when interpreting the social world (Keysar, Lin & Barr, 2003). Past work has shown that motivations (manipulated by financial incentives) influences the extent to which mentalizing judgments are biased toward one's perspective (Epley et al., 2004).

The present work extends this idea to socially-derived motivations. Intentional harm usually produces blame motivation (Alicke, 1992; 2000, Darley & Pitman, 2003). And motivations tip judgments in directions that are consistent with those motivations (Kunda 1990, Ditto, Pizarro & Tannenbaum, 2009). Structurally then, influence can flow from perceived intent to perceivers' motivations, and from perceivers' motivations to their judgments. This conceptual description aligns well with a statistical (meditational) description of the observed data (across

two studies and one replication). Moreover, effects were consistent regardless of whether the person enacting the harm was a friend (Study 1), an object of romantic desire (Study 2), or a third party (Study 3), and regardless of whether the paradigm was an adaptation of a classic psychology task (Study 1), a short story focusing on one protagonist's internal mental life (Study 2), or a humorous video (Study 3). Also, despite these methodological differences, the magnitude of the observed effect was stable across the three studies, including the two replications ($d_s = .40-.44$).

Intent Oversensitivity, or Failure to Track Beliefs?

A possible alternative explanation for these effects is that participants were not influenced by motivation to blame, but rather, failed to represent the victim's beliefs separately from the participants' own (more informed) beliefs. However, as reviewed in the introduction, decades of research on belief attribution and intentionality tracking suggest that this is unlikely. Moreover, we independently verified participants' ability to separate their own beliefs from the protagonist's beliefs in Study 1, and directly gave participants the relevant second-order belief in Studies 2 and 3. In addition Study 2 showed that the inclusion or exclusion of the few participants who incorrectly assessed the victim's beliefs had no substantive impact on our results. This again highlights the distinction between what information people have, and what information they actually deploy in the service of social cognition (Keysar et al., 2003).

Functional Advantages of Intent Oversensitivity

Using structurally irrelevant information to inform social inferences is not, strictly speaking, sensible. In the case of affective mentalizing, this tendency threatens empathic accuracy. Empathic accuracy, while rarely perfect, is nonetheless an important component of adaptive social functioning, and has been throughout human evolutionary history (Zaki, Weber,

Bolger & Ochsner, 2009). Normally, people's mentalizing toolkit helps them to construct useful barriers that block such threats. Why then are people oversensitive to harmful intentions when inferring others' emotions? Or, to pose a more general question, how would such oversensitivity be psychologically functional? In Chapter 1, we suggested two answers to this question: first, we noted that such responses are psychologically functional to the individual. Motivations produce biases that facilitate the satisfaction of those motivations (Kunda, 1990; Balci et al., 2006). Intentional harm produces motivation to blame, and more harm provides a stronger basis for blame: It is easier to (satisfy the motivation to) blame, condemn, and punish someone who caused a lot of harm rather than a little. The righteous satisfaction that one receives may also be greater for punishing large harms, rather than minor transgressions. This idea seems equally applicable to emotional harms as to the financial harms described in the previous (and next) chapter.

Second, we noted that because human beings evolved (and still live) under conditions of extreme interdependence (Kurzban & Leary, 2005), it is functional for people to react strongly to intentional harm (Hoffman et al., 1998; Gollwitzer & Rothmund, 2009). Oversensitivity to intentional harms protects social groups from socially malfeasant individuals, causing such individuals to be detected more easily and dealt with more decisively. This work suggests, however, that it may also interfere with empathic accuracy (or other harm estimates) under motivated circumstances. However this bias may be heuristically functional insofar as intentional harms usually *are* worse than unintentional harms, since they are expressly tailored to produce harm. "Intentional harms are worse" may be a particularly useful heuristic when navigating the highly complex decision space of empathic accuracy (Zaki & Ochsner, 2012).

Motivations Affect Judgments

Empathic judgments are complex and subjective—precisely the conditions under which motivational biases tend to be largest (Ditto & Boardman, 1995; Dunning, Leuenberger & Sherman, 1995). However, the previous chapter suggested that blame motivation can inflate harm judgments even when the harms in question are quantified objectively (e.g., in dollars and cents). The fact that this bias emerges even when people are financially incentivized to be accurate suggests that it is not merely expressive—rather, in line with previous research on motivated reasoning, people maintain an illusion of objectivity concerning their own judgments (Kunda, 1990; Pyszczynski & Greenberg, 1978), allowing motivated reasoning to subtly affect genuine beliefs (Ditto, Pizarro & Tannenbaum, 2009). Of course, not all harms can or should be reduced to financial losses, or other objective metrics. In fact, objective harms are only harmful to the extent that they impact subjective experiences. These subjective experiences were examined more directly in the present chapter.

Childish Mentalizing in Adults?

The present work demonstrates that intents that are structurally irrelevant to mental state inferences can nonetheless impact those inferences—even among adults who demonstrably possess mental state reasoning tools that should prevent this (including false belief inference, mental state bootstrapping, and second-order mental inference; Harris et al., 1989; Miller, 2009; Wellman et al., 2001). As such, the present studies appear to be in tension with a healthy corpus of research on developmental psychology and mental state reasoning. However, our suggestion certainly is not that the broad conclusions of these literatures are incorrect. We would, however, suggest that the psychological processes being studied are likely to be affected by the social motivations of the perceiver, and current work in this area would benefit from taking perceivers' motivations into account (preferably by varying them systematically). While much is sometimes

made of the precise point in human development at which various mentalizing abilities begin to function properly, their proper function, at any age, may depend on non-developmental factors, (Baillargeon et al, 2010, Sodian 2011), including the motivations of the perceiver.

In fact, the studies reported here are far from the first to show that apparently childish mentalizing is sometimes observed in adults (Birch & Bloom, 2007; Lin et al., 2010; Newton & de Villiers; 2007). For example, when the standard false belief task is made significantly more cognitively demanding, and when more sensitive measurement instruments are used, adults will sometimes demonstrate deficits in false belief performance that are similar to (though smaller than) those deficits commonly observed in young children (Birch & Bloom, 2007). Somewhat contrary to present findings however, these signs of false belief confusion in adults emerged only when the positions of the relevant containers (the hidden object locations in a more demanding version of the false-belief task) were “scrambled” after the participant left the room, with the additional constraint that the missing object ended up in its original physical location (but in a different container). The protagonist (“Sally”) might plausibly look in the original location without taking note that the container had changed; under this interpretation, the previous findings do not necessarily indicate belief confusion. There would also seem to be another alternative interpretation of this previous work: Sally might believe that there was some chance (participants gave probabilistic responses) that whomever moved the containers (“Anne”) would have left the object in the same physical location, either because it belonged there, or to make it easier to find. Thus at least two relevant beliefs may have been at play. The current research extends this idea, demonstrating that even without the set of restrictions described above, adults may still fail to discount their own privileged knowledge in making their judgments (a curse of knowledge effect). If there appears to be a discrepancy here between Birch & Bloom (2007) and

the present work, one possible explanation may be that reasoning about beliefs pertaining to physical locations is somewhat different from reasoning about beliefs pertaining to others' minds. It seems plausible that people are more sensitive to the bad intentions of social agents than to the locations of objects (Gollwitzer & Rothmund, 2009)—which might make knowledge of such intentions a particularly powerful curse, and one that may manifest itself across a somewhat wider range of situations.

Strengths, Limitations and Future Work

As noted briefly in the introduction, the work reported here required that participants simultaneously mentalize about intentions, beliefs, and emotions. In the real world, successful mentalizing often demands this sort of cognition: thinking about multiple mental states at once (e.g., Baron-Cohen, 1991). Yet researchers almost always attempt to isolate mental inferences from one another—carving mentalizing at its joints. While this approach is clearly invaluable, it may also be worthwhile to conduct some studies that reflect, in a simplified way, the combinatorial realities that often define mental state inferences. After all, people's intentions partly depend on their beliefs; their beliefs are partly a function of their emotions; and their emotions depend partly on their intentions and whether or not those intentions are fulfilled—as well as what other people intend and feel toward them, and what they believe about them (Malle & Holbrook, 2012; Wellman et al., 2001).

It is encouraging that the statistical mediation models described here consistently comport well with the proposed theory (but poorly with alternative models), and that the effect is robust to different kinds of stimuli. Still, the subjective nature of emotional harm makes interpretation of the observed effects more difficult than the interpretation of the objective harm estimates observed in Chapter 1. Notably, it is much less clear what 70 “units of emotional harm”

means, than “\$4,862 in financial losses”). While emotional inference is inescapably subjective, a wider range of stimulus materials would provide additional inferential leverage. Future work might focus on affective inferences from pictures or videos of real human faces. While the cognitive and neural mechanisms underlying social inferences based on faces are different from those underlying non-face-based inferences (e.g., Ames, Fiske & Todorov, 2011; Jenkins & Mitchell, 2011; Van Overwalle, 2009), the effects of motivation on these two types of inferences may be functionally similar (or not).

A broader limitation of this chapter is that it provides an argument for the general construct of “motivated mentalizing,” but focuses disproportionately on mentalizing about feelings (albeit feelings that depend on beliefs and intentions), and the specific motivations that arise in response to intentional harm. Motivated perceivers might well misattribute thoughts, desires, and other transient mental states because of many different motivational forces—for example (returning to faces), people probably do sometimes prefer certain interpretations of their romantic partners’ ambiguous facial expressions; in these cases perceivers’ motivations (whether driven by optimistic or pessimistic forces) might alter one’s judgments of one’s partner’s affect. As another example, the motivation to belong or to feel understood might lead people to over-attribute shared in-the-moment thoughts to desirable others or to one’s own ingroup.

Some existing evidence is consistent with these ideas—however an understanding of the general psychological processes (if any) by which motivation affects the inference of transient mental states seems greatly to be desired. In this vein, we would also highlight again the inherent limitations of mediation analyses such as those presented here. These are a starting point for understanding psychological mechanisms—not a conclusion in themselves. Chapter 3 attempts to deal directly with the question of whether participants are actually *motivated* to blame.

Calibrating Responses to Harm

Beliefs are, in some sense, supposed to be true. This is what beliefs are for: representing reality (Leslie, 1994). However, beliefs sometimes *misrepresent* reality. Therefore, the realization that people live in a world made up of thoughts and ideas as much as of objects and events is a necessary precursor for understanding human behavior. In the case of affective mentalizing, this mind/world distinction is critical for adaptive social responding. Consoling a friend who feels especially bad is one of the most important things that a friend can do. Consoling a friend who does not need or want consoling is inappropriate and socially awkward. In addition, tuning one's responses to harms of all kinds (both subjective and objective) in proportion with the true magnitude of those harms is important because the resources that people use to respond to harm (time, money, emotional energy) are limited. Taken in conjunction with work presented in Chapter 1, the present chapter suggests that one factor that may be inappropriately influencing this tuning is people's oversensitivity to intentional harms. We next turn to the question of whether these effects truly include a motivational component.

CHAPTER 3: BUT IS IT REALLY MOTIVATED?

Chapter Summary

The previous two chapters showed that people overestimate the damage caused by intentional harms. This chapter further examines why this occurs, specifically targeting the motivation to blame the intentional harmdoer. Study 1 rules out an alternative explanation for the earlier result in terms of human versus natural causes, showing that the harm-magnification effect emerges specifically for human intentional action, but not for human unintentional action. Study 2 asks whether intended harm motivates blame or merely demonstrates the actor's intrinsic blameworthiness. Consistent with a motivational interpretation, participants freely chose blaming, condemning, and punishing over other appealing tasks in an intentional harm condition compared with a control condition. Study 3 also measures motivation, but with converging indicators of persistence (effort, rate, and duration) in blaming. Together, these studies argue that intentional harms are worse, even when they are not, at least partly because people actively want to blame the intentional harmdoer. Perceived intent emerges as catalyzing a motivated social cognitive process related to social prediction and control.

Introduction

The previous two chapters suggest that people accord extra weight to the consequences of bad intent. Earlier work (e.g., Alicke, 2000; Alicke & Davis, 1990; Robinson & Darley, 1995) and statistical mediation analyses presented earlier suggest blame motivation as a likely psychological mediator of this effect—with intentional harms leading to blame motivation and blame motivation, in turn, leading to exaggerated harm judgments. However it is difficult to say with certainty whether participants in these studies were truly *motivated* to blame—not least of all because self-report measures of participants’ desire to blame, punish, and condemn could be interpreted as judgments of *blameworthiness* rather than *blame motivation*.⁹ After all, people do find intentional harmdoing more blameworthy than unintentional harm (e.g., Alicke, 1992; Darley & Pittman, 2003; Young & Saxe, 2009). Arbitrating between motivated accounts of biased judgment and “purely cognitive” accounts of such judgments is notoriously difficult (Kunda, 1990); however one promising approach to assaying motivation is to examine what participants choose to do of their own free will, and how hard they work in the absence of any external incentive. The present chapter attempts to apply this approach. Study 1 aims to rule out an alternative explanation for previous findings, setting the stage for a more principled investigation of mechanism. Study 2 examines participants’ free choice behavior as an index of their motivations. Study 3 measures participants’ motivation by indexing persistence.

Participants

⁹ It is unclear, however, why or how a mere assessment of someone’s blameworthiness could function as a psychological mediator between perceived intent and perceived harm.

Across the three studies, 630 participants were recruited using Amazon Mechanical Turk and then redirected to an experimental Web site (Experiment 1: $N = 309$, $M_{\text{age}} = 30\text{y}$ (10.5), 121 male, 187 female, 1 not reported; Experiment 2: $N = 201$, $M_{\text{age}} = 34\text{y}$ (13.5), 1 not reported, 90 male, 111 female; Experiment 3: $N = 120$, $M_{\text{age}} = 33\text{y}$ (12.2), 51 male, 68 female, 1 prefer not to say). Participants were paid standard rates in exchange for their participation (Experiment 1: \$0.40, \$6.26/hr; Experiment 2, \$0.40, \$5.39/hr; Experiment 3: \$0.40, \$5.04/hr. Pay rate for Study 3 is calculated from Amazon Mechanical Turk acceptance/submission time, rather than actual task start/end time, and is therefore likely an underestimate). Recruitment was limited to the United States, and to persons with at least a 95% approval rating in their previous work. See Chapters 1 and 2 for research on the reliability and representativeness of this sample. Participants who had completed any previous study in this line of work were automatically prevented from accessing the recruitment page. For Studies 2 and 3, all participants were included in all analyses. Study 1 included basic data quality checks. For complete information on these checks and the number of persons excluded by each one, see Appendix G).

Study 1

Before we directly address the mechanism underlying the magnification of intentional harms, we wish to rule out a possible alternative interpretation of the “objective harm” studies from Chapter 1. In these experiments, intentionality was manipulated by comparing an intentional, person-caused harm with an unintentional nature-caused harm. Thus, the comparison of intentionally caused harms vs. unintentionally caused harms was perfectly confounded with a comparison of person-caused harms vs. nature-caused harms. The person/nature-caused comparison in these experiments was made deliberately, with an eye toward policy-related questions about resource

allocation (discussed in Chapter 1). However it cannot be said definitively from these experiments whether the effects were due to intentionality per se: people may overestimate human-caused harms, regardless of intent—perhaps because of a biased belief that human-caused harms do tend to be larger. Admittedly, it is difficult to explain such a bias would exist without making some reference to intent (e.g., “humans can intentionally calibrate their actions to cause harm, whereas nature cannot”), however, it seems worth addressing the point empirically.

Experiment 1 sought to investigate a harm that was person-caused, while varying only intent. If the “person” hypothesis is correct, we should expect inflated estimates, relative to ground truth, in both conditions. If however the “intentionality” hypothesis is correct, we would expect accurate estimation of unintentional harm, but inflated estimates of intentional harm.

The methodological approach here was essentially identical to that employed in Studies 4 and 5 of Chapter 1: Participants read a vignette describing a harmful action, saw a list of all the harms arising from that action (in dollars), and then estimated their sum. While the set of numbers (different from those used in Chapter 1) was identical across conditions, the vignette framed these costs as accruing from either an intentional harm or an unintentional harm (randomly assigned).

Methods

The framing vignette described a nursing home employee who mixed up patients’ medications (see Appendix H for full text). This vignette was a composite based on various academic papers, government reports, and news releases concerning high rates of inappropriate medication in nursing homes (e.g., giving antipsychotics without any clinical indication for this treatment, ignoring maximum recommended daily doses; Centers for Medicare & Medicaid

Services, 2013; Chen et al. 2010; Span, 2010). From these reports, which included instances of both intentional and unintentional giving of inappropriate medication, we constructed a composite narrative, in which inappropriate medication was dispensed either intentionally or unintentionally (50/50 random assignment). Across conditions, the vignette featured the same act (mixing up the pills), committed by the same person (an employee named Jake), resulting in the same harms (symptoms and associated medical costs), to the same people (nursing home residents). The proximal cause of the to-be-estimated harms was also the same across conditions (inappropriate medication). Only Jake's *intent* varied between conditions. Participants were fully informed that this vignette was a composite, rather than a veridical report of a specific instance: "The following story is a composite of several true events. We are not asking about any specific instance. Resemblance to any specific person(s) or location(s) is coincidental."

After reading the composite vignette, participants were informed that they would next see all of the harms arising from the actions described in the vignette, along with the associated medical costs (in dollars). They were further told that this information would go by quickly, and that they should pay close attention, as they would be asked to report these costs. Participants then viewed eight consequences of the medication mix-up, along with their associated costs (e.g., "overtime staff hours: \$271.10." The items were presented for 2000ms each, and totaled \$4,433.43. After all eight items had been presented, participants were asked to "Estimate the TOTAL cost (i.e., the sum of the numbers you just saw)." Participants were given 15 seconds in which to make this estimate (more than in the previous studies). After making their estimates, participants completed a manipulation check as to the intentionality of the harm: "Did Jake give the wrong medications on purpose, or on accident?" A reading check asked "Did all of the residents recover in the end?" (see Appendix G for participant exclusions). Participants then had

the option of giving their opinions as to whether Jake should lose his job (60% “yes,” 38% “no,” 2% “not sure/no opinion”), and whether it would be better or worse for the residents to know what caused the uptick in medical issues, assuming that no such mix-up ever occurred again (51% “they’d be” better off knowing, 40% “they’d be better off not knowing,” 9% “not sure/opinion”). Finally, they reported their demographic information.

Results

As in Chapter 1, participants’ estimates were accurate when the harm was framed as accidental (\$4,557.20, $SD = 2,261.70$, vs. the correct answer of \$4,433.43, $t(117) = 0.59$, $p = .55$). However, participants who interpreted the same set of numbers as arising from intentional harmdoing significantly overestimated their sum ($M = \$5,224.17$, $SD = 3,903.22$, $t(130) = 2.32$, $p = .02$, $d = 0.20$), relative to ground truth. The two experimental conditions differed marginally from one another, $t(212.2) = 1.67$, $p = .097$, adjusted for unequal variances. This finding aligns well with our earlier interpretation of these findings, but comports poorly with the idea that the results observed in Chapter 1 were driven by the person-caused/nature-caused difference.

Study 2

We have suggested blame motivation as a possible psychological mediator of this effect, and in fact, motivations have been found to produce objective magnitude biases in previous research (e.g. Balci et al., 2006). Statistical mediation analyses in the previous two chapters are consistent with this idea. However, recent work has reaffirmed the importance of treating statistical mediation as a starting point for determining causality, rather than as conclusive in itself (e.g., Bullock, Green & Ha, 2010; Imai, Keele & Tingley, 2010; Fielder, Schott & Meiser,

2011). In previous experiments, we asked participants to assign degrees of blame, moral condemnation, or punishment to offenders. Here, rather than asking participants to make these ratings, we simply presented them with the opportunity to do so. Our reasoning was that, if participants were motivated to blame, condemn, and punish, they would choose to do so, perhaps even when an otherwise more appealing task was available.

Methods

To link these findings to a situation known to produce the harm magnification effect, participants read one of the two versions of the same nursing home vignette as in Study 1 (randomly assigned). After reading the vignette, participants answered three questions before proceeding to the main task: “Did Jake give the wrong medications on purpose, or on accident?” (manipulation check); “Do you feel like you remember hearing/reading about an event like this (e.g., on the news)?” (5% “yes,” 26% “maybe,” 69% “no”); and “Assume that Jake never mixed up pills again, and that there was no long-term harm done. Do you think that the residents of the nursing home would be better off knowing why they got sick, or not knowing why they got sick?” (62% “they’d be” better off knowing, 28% “they’d be better off not knowing,” 10% “not sure/opinion”). Seven participants (3%) answered the manipulation check incorrectly. Their inclusion or exclusion made no difference, and they are retained in the analyses reported below.

Next, participants completed the main task, and then reported their demographic information (age, gender, and race). In the main task, participants saw a page with the following instructions: “Thanks—you’re about 75% done. Please choose one of the following to complete (pending availability). They all take about the same amount of time.” Participants chose freely among the following five options, presented in a random order:

Offer your opinion about how (if at all) someone like Jake should be punished for what he did, and what amount of blame/moral condemnation he deserves

Take a short quiz about the cost of healthcare in the U.S. (answers will be provided at the end of the quiz)

View a healthcare advertisement and offer your opinions about its tactics and effectiveness

Offer your opinion about elder care in the U.S.

Answer various questions about nursing home situations like one you just read about

Participants selected one task to complete. Selection of the “blame/punish/condemn” option (presented first above, but again, presented in random order to participants) was the critical index of motivation to blame, punish, and condemn. The other four options controlled for various alternative explanations. For example, participants’ tendency to select the “blame/punish/condemn” option more in the *intentional harm* condition than in the *unintentional harm* condition could, in principle, occur because the *intentional harm* condition made participants more concerned about eldercare, and caused people to want to express their feelings about this—but without actually generating any motivation to blame. This interpretation would predict an increase in selections of both the “blame/punish/condemn” option *and* the “opinions

about elder care” option (a pattern that was not observed, see Results). Another possibility might be that reading a vignette about a healthcare-related situation could make the concept of healthcare accessible, thereby increasing processing fluency for tasks like the “blame/punish/condemn” one above, thus increasing participants’ tendency to select this option; however, because all five options relate to healthcare, it is unclear how this explanation would predict a unique increase in choices to blame/condemn/punish.

Results

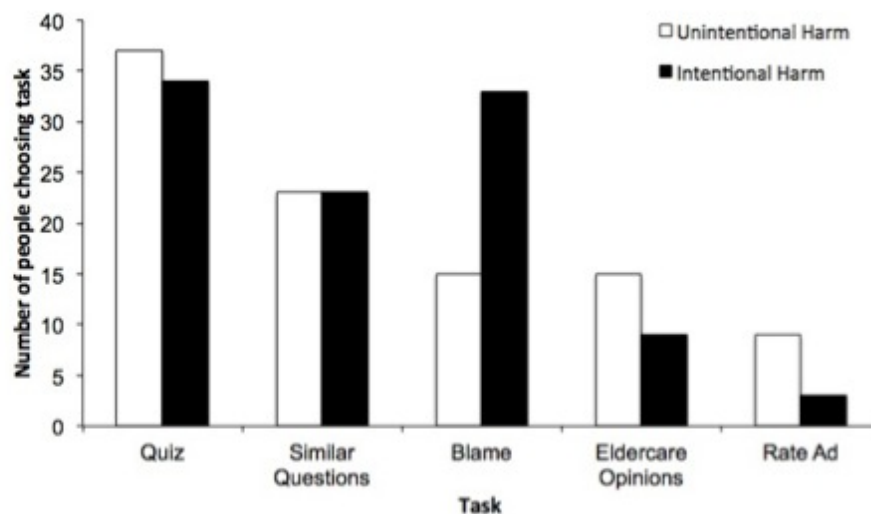


Figure 1. Number of participants selecting each task, separated by condition.

A unique increase in choices to blame/condemn/punish was, in fact, observed (Fig. 1). In the unintentional harm condition, the blame/condemn/punish task was tied for third-most popular among the five options, while the option to take a short quiz was by far the most frequently selected. However, in the *intentional* harm condition—that is, when the vignette was identical except that the medications were switched intentionally, selection of the “blame” task more than doubled, $\chi^2(1) = 6.7, p = 0.009$. As discussed in the Methods section, free selection of any of the other tasks could have been affected as well—for example, by increasing the cognitive

accessibility (and thus fluency-based selection) or interestingness of various healthcare-related issues. In fact, however, preferences for all other activities were unaffected by the intentionality status of the harm. This unique sensitivity of the *blame* task was statistically reliable regardless of whether or not a Bonferroni correction was used to evaluate participants' preference for each of the five activities ($\alpha=.05$ or $\alpha=.01$).

This pattern was corroborated in an overall chi-square analysis of all intent/choice combinations, $X^2(4) = 11.4, p=0.02$. This result reflects the fact that the blame task was chosen numerically more often in the intentional condition, while all other tasks showed either the opposite pattern, or no difference across conditions. An alternative analytical approach is to collapse across the four control conditions and compare the controls against the critical condition in a 2x2 chi-square analysis. This approach again yields a significant effect of the harmdoer's intent on participants' task selections, $\chi^2(1) = 8.2, p = 0.004$.

It is somewhat difficult to make sense of participants freely *choosing* to do something unless we suppose that they *wanted* to do that thing. The concept of “wanting” is inextricably linked to motivation. These data therefore provide some support for the notion that intentional harm-doing elicits motivation to blame, punish, and condemn the harmdoer—making participants more likely to enact these behaviors of their own free will, without any external incentive, even when otherwise more appealing activities are available.

Study 3

Motivation is sometimes expressed, not only in what people choose to do, but also by the intensity or frequency with which they do it. As the difficulty of a task increases, more effort is required. Individuals may then give up (suggesting relatively low task motivation), or they may

raise their level of effort to match the increasing demands (suggesting relatively high motivation). Following this logic, participants in Study 3 completed a reinforcement game modeled after classic psychology studies of animal motivation (e.g., Ferster & Skinner, 1957; Lattal & Gleeson 1990; see also Hursh et al. 1988; see DeGrandpre, Bickel, Hughes, Layng, & Badger, 1993 regarding the statistical logic of this approach). In a typical experiment of this type, the number of times that a rat pressed a lever was taken to index the rat's motivation to obtain the food reward associated with that lever (such that hungry rats would demonstrate greater effort and persistence in lever-pressing than non-hungry rats). Our Study 3 similarly indexed effort and persistence in humans—measuring mouse clicks rather than rat presses—to examine motivational responses to intentional and unintentional harms (see also Morewedge, Huh, & Vosgerau, 2010).

Methods

After reading the same vignette as in Studies 1 and 2, participants answered two questions before proceeding to the reinforcement game (main task): “Do you feel like you remember hearing/reading about an event like this (e.g., on the news)?” (4% “yes,” 7% “maybe,” and 89% “no”) and “Did Jake give the wrong medications on purpose, or on accident?” (manipulation check). Four participants (3%) answered the manipulation check incorrectly. Their inclusion or exclusion was negligible, and their data were retained in the analyses reported below.

After answering the familiarity and manipulation check questions, participants proceeded to the main task. At the top of the screen was written “Your opinion: Should Jake be investigated? Punished?” Also onscreen were task instructions, a picture of a petition (with twenty blank spaces for signatures), and a picture of a blue envelope. The petition called for the nursing home employee featured in the vignette to be investigated, and for appropriate actions to

be taken. Above the envelope was written “Click envelope repeatedly to try to get signatures.”

The task instructions read as follows:

If you think that Jake should receive some blame, punishment, and/or moral responsibility for what happened, please indicate this this by clicking on the blue envelope below. Each time you click on the envelope, you will "send a letter" to recruit a signature for the (pretend) petition, signaling that Jake should be investigated and appropriate actions should be taken.

Just like in real life, not every letter will get a response. The more signatures you get, the more strongly we assume you feel that Jake should be investigated, blamed, punished and/or held morally responsible.

Participants could click as many or as few times as they wished (or not at all) to send increasing numbers of letters. Participants knew that the situation described in the vignette was a hypothetical composite, and that the petition in the reinforcement game was only pretend. As noted above, the instructions explained that not every letter sent would elicit a response. This feature of the design allowed for reward schedule variation. The reinforcement schedule changed over time, such that progressively more work (more clicking) was required to earn each signature. The game began with a fixed-ratio reinforcement schedule of 3 (FR3), in which one out of every three clicks produced a signature. Which of the three clicks in each triad triggered the signature was randomized (i.e., in clicks 1-3, the third click might produce the signature; in clicks 4-6, the fifth might produce the signature). Every time participants earned five signatures, the reinforcement ratio increased by three, progressing through FR6, FR9, FR12, FR15, FR18, FR

21 and FR24. Participants were free to end the game at any time, and knew that they would earn no more or less money as a function of how long they spent clicking.

While “number of signatures earned” could be interpreted as a Likert-type scale (more signatures equating to higher blame ratings), this interpretation ignores the fact that participants had to do work to earn each signature. In keeping with other research in reinforcement learning (DeGrandpre et al., 1993; Hursh et al., 1998), we reasoned that doing more work to earn a reward (in this case, a signature) indicated greater motivation to earn that reward.

Following the main task, participants were presented with two optional questions: whether they had any opinion about what should happen to the person who mixed up the medication, and whether they thought it would be better or worse for the residents to know what caused the sudden uptick in medical issues, assuming that the incident was never repeated. Finally, demographic information was collected.

Results

Participants did substantially more work in the intentional harm condition ($M = 174.2$ clicks) than in the unintentional harm condition ($M = 25.6$ clicks), $t(67.59) = 6.0, p = 8.7 \times 10^{-8}$, corrected for unequal variances. This increase in effort was observed no matter how effortfully demanding the task became (that is, across all eight reward schedules examined). These results remained stable if counts were log-transformed ($M_{\text{intentional}} = 1.88, M_{\text{unintentional}} = 0.72, t(114.28) = 8.1, p = 6.3 \times 10^{-13}$), or when time spent clicking, rather than number of clicks, was used as the unit of analysis ($M_{\text{intentional}} = 1:21, M_{\text{unintentional}} = 0:51, t(84.12) = 3.4, p = .001$). Mean click rate was also significantly greater in the intentional harm condition ($M_{\text{intentional}} = 2.0$ clicks/second, $M_{\text{unintentional}} = 0.5$ clicks/second, $t(84.85) = 7.4, p = 1.0 \times 10^{-10}$). The details of these calculations are reported in Appendix I. In sum, intentional, rather than unintentional harm caused participants to

work more, to work faster, and to work for a longer period of time. When people work more, faster and longer, in the absence of any external incentive to do so, it seems fair to suggest that they are motivated.

Response elasticities (number of responses per unit price; DeGrandpre, et al., 1993) were also computed to assess the relationship between effort and reward schedule (Fig. 2). A 2 (Condition: intentional, unintentional) $\times$ 8 (Reinforcement schedule: FR3-FR24) mixed ANOVA, with condition as a between-subjects factor and reinforcement schedules as a within-subjects factor, revealed a main effect of intent, with the intentional harm condition eliciting more effort and persistence than the unintentional harm condition, $F(1, 101) = 60.76, p = 4.5 \times 10^{-12}$, partial $\eta^2 = .374$. Consistent with previous research, a main effect of reinforcement schedule also emerged, confirming that response elasticities decreased as demandingness of the task increased $F(2.8, 40.2) = 63.96, p = 2.5 \times 10^{-30}$, partial $\eta^2 = .362$, Huynh-Feldt-corrected for nonsphericity. There was also a significant condition $\times$ reinforcement schedule interaction, $F(2.8, 300.89) = 15.89, p = 4 \times 10^{-9}$, partial $\eta^2 = .129$, Huynh-Feldt-corrected for nonsphericity, suggesting that *unmotivated* participants reduced their effort linearly as signatures became harder to obtain, while the *motivated* participants initially raised their level of effort to match the increasing task demands, then eventually decreased their effort (though not to the level of the unmotivated participants) as task demands became extreme. In summary, when the going gets tough, the motivated keep going, and, by this standard, intentional harm elicits motivation to blame.

General Discussion

The studies reported in this chapter go beyond our prior meditational evidence, which is only correlational, to experiments measuring motivation in terms of choice and persistence

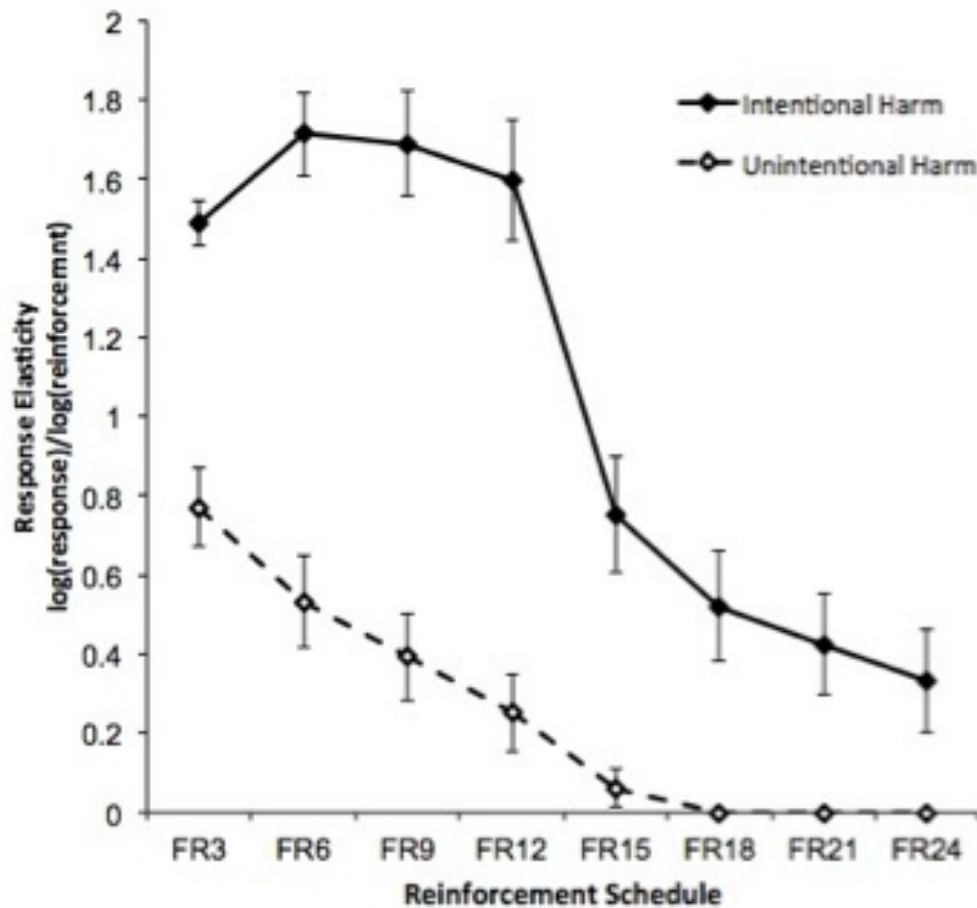


Figure 2. Motivation to earn signatures for mouse clicks (expressed as response elasticity) as a function of intentionality condition (intentional vs. unintentional harm). Horizontal axis shows fixed-ratio reinforcement schedule (for example, in FR3, every third response was reinforced). Error bars represent standard errors.

(effort, rate, and duration). These studies also focus more precisely on human intended and unintended harm. Results support the suggestion that intentional human harmdoers motivate observers' blame, causing people to overestimate the damage done. However future research

must further interrogate the proposed causal relationship between this blame motivation and harm magnification.

SUMMARY AND APPLICATIONS

Summary of Empirical Findings

Together, these studies suggest that intentional harms are worse, even when they are not, and that one likely cause of this (there may be several) is that people actively want to blame people who hurt each other intentionally. Chapter 1 revealed that this effect is observed when participants themselves decide what harm means (Studies 1-3), and also when harms are quantified in dollars and cents (Study 4). Moreover, the effect was robust to financial incentives for accuracy, and to the explicit separation of “compensatory and punitive damages” (Study 5). Chapter 2 extended this idea into the social-cognitive domain. Study 1 showed that participants could pass a false belief task and then, moments later, fail an emotional inference task by failing to apply that belief. We suggest that this occurred because motivation to blame intentional harmdoers led to biased processing strategies (as they often do, Kunda, 1990), with the result being that perceptions of harm were inappropriately inflated. Identical effects (and effect sizes) were observed regardless of whether the victim believed the harm to be intentional (Study 2, plus replication) or unintentional (Study 3, plus replication), or when no explicit belief statement was given (Study 1), and regardless of whether the task was formatted as a classic social cognition task (Study 1), a short story (Study 2) or a humorous video (Study 3). Meditational results (and model comparisons) were highly consistent across Chapters 1 and 2, despite differences in stimuli and the type of harm under examination. Chapter 3 sought to go beyond these meditational models to ask experimentally whether participants were truly motivated to blame. Study 1 ruled out an alternative interpretation of earlier results, and replicated the magnification of objective harms with new stimulus materials. Study 2 showed that intentional harm leads people to actively choose to blame, condemn and punish, even when a more desirable task is available (motivation as

choice). Study 3 showed that intentional harm leads people to work longer, harder and faster to blame, condemn and punish (motivation as persistence). Taken together, these studies provide evidence that the motivation to blame intentional harmdoers can lead to magnified perceptions of intentional harm.

Applications

We have said relatively little about the broader implications of this work; however the magnification of intentional harms would seem to be of pragmatic importance from a legal and policy perspective. First, as noted briefly in Chapter 1, policymakers must deploy limited resources to rectify (or prevent) harms to society (McGraw et al., 2011). Of critical importance to this process is a reasoned estimation of the magnitude of these harms. We suggest that this estimation may be systematically biased upward if the harm in question is perceived as intentional. Second, the concept of intentionality is of deep importance to the law, and determining monetary compensation for both intentional and unintentional harms is an important part of the law's function. The effects revealed here raise new questions about how the law fulfills this function. We examine these two sets of issues (budgetary policy and legal implications) in turn.

Budgetary policy

Policymakers and legislators make consequential decisions about whether or not to intervene when harm-causing events take place. The present work suggests that power-holders may be biased toward engaging in (often extremely expensive) interventions when there is a “bad guy” to fight—that is someone intending the harms being brought about, who elicits the motivational forces we have described. Yet it may be that more good could be done at much less cost elsewhere, where there may not be any intentional agent responsible. This idea is supported

by the findings presented toward the end of Chapter 1, wherein participants overestimated, by a considerable margin, the damage done by a minor act of environmental terrorism. While it may seem sensible to do “everything we can” to prevent terrorist acts, a full commitment to stopping one terrorist act necessitates de-commitment from other problems.

For example, according to the most recent published estimate from the Congressional Budget Office (CBO report; 2010), the Iraq War will cost about \$1.9 trillion by 2017 (about \$125 billion annually from 2003-2017). It could be argued that, whatever the cost, this military action constitutes a moral imperative. However, regardless of whether it is a moral imperative, it is certainly (also) a pragmatic decision involving painful tradeoffs.

These tradeoffs are made clear by considering the cost-benefit analysis of fighting global vitamin A deficiency. Micronutrient deficiency is a social ill that lacks an intentional antecedent; yet it is responsible for more than a million deaths per year (World Food Program Annual Report, 2007). This is a large-scale problem with a simple solution: provide vitamin A capsules. This practice is demonstrably effective in preventing death and suffering. A USAID Vitamin A Supplementation Program conducted in Ghana, Nepal, and Zambia prevented about 44,000 deaths per year, with a comparatively modest annual program budget of \$2.9 million (USAID Micronutrient Program, 2004). These efforts are ongoing, but at present, the scale of the intervention is far less than the scale of the problem, and more investment would save more lives.

Speaking strictly in terms of lives saved per dollar spent, in order to match the Vitamin A program’s average annual efficiency of \$65.00 per life, Saddam Hussein would have needed to cause about 2 billion deaths every year for 15 consecutive years to justify the \$1.9 trillion CBO estimate. This is, of course, impossible: the world’s current population is only approximately 7 billion. Even this decidedly lopsided comparison is likely an underestimate of the economic

incongruity. The analyses of the Copenhagen Consensus (Horton, Begin, Greig, & Lakshman, 2008) suggest that a combined annual investment of \$60 million in vitamin A and zinc supplements would yield returns of more than \$1 billion per year in productivity and other benefits, with each dollar spent generating more than \$17 in returns.

As satisfying (and indeed, beneficial) as it may be to effect political change in unstable regions, or to respond decisively to terrorist attacks, policymakers cannot afford to be insensitive to the math. The stakes are too high. As noted in Chapter 1, there are many good reasons why policymakers might invest more resources in combatting terrorism than in fighting vitamin A deficiency (e.g., the perceived preventability of future harms). However, there can be little doubt that moral sentiments and political pressures sometimes influence important policy decisions in ways that lead to unsettling consequences. The present work suggests that, even when policymakers are attempting to make purely (or partly) consequence-based policy decisions, those decisions may carry an element of moral bias. Indeed, the data serving as the *inputs* of those decisions may already carry this bias. Budgetary proposals contain detailed mathematical analyses of costs and benefits; yet given the scale and uncertainty of major government actions, many of the figures must be estimated (what were the precise costs to tourism associated with 9/11?). The present work suggests that estimates, even when made in good faith (and under accuracy incentives) could be biased upward by the presence of harmful intent. Naturally, the public sector officials who generate these estimates did not face precisely the same demands as those that were present in our experiments. Yet government workers face time pressure, distraction, and other forms of cognitive load, and the data they analyze are often orders of magnitude more complicated than the simple arithmetic posed to our participants. Thus, similar principles may apply.

Legal considerations

This work aligns with theoretical perspectives advancing the need for the law to align with the cognitive realities of human psychology (e.g., Darley & Huff, 1990). The civil legal system is predicated on the idea of making victims “whole.” To accomplish this, jurors are often asked to separately define two quantities: *compensatory damages* and *punitive damages*.

Compensatory damages refer to the sum of money equal to the real financial losses of the victim (e.g., hospital bills, property damage). Punitive damages refer to additional sums that seem to constitute appropriate punishment (usually paid by the harmdoer to the victim), for example, due to particularly malicious intent. Estimates of compensatory and punitive damages were made by participants in Study 5 of Chapter 1, where they produced two separate figures: (1) the “...total of the damages, in dollars (the sum of the numbers you just saw),” and (2) “extra money ...that person [should] have to pay in fines/penalties/compensation beyond the actual costs as shown (that is, beyond the sum of the numbers you just saw).” The two calculations are explicitly separated by the law; however, the present research suggests that this separation may not be psychologically plausible. Jurors faced with a case involving intentionally-caused damages will likely overestimate the required compensatory damages (even when accurate cost information is available, and even when the jurors are motivated to be accurate). This inaccuracy is compounded if punitive damages are added for malicious intent, effectively punishing the defendant twice for the same thing.

It may in fact be worth considering whether a “litigious society” mentality can partly explain why intent magnifies harm. It seems likely that when people encounter harmful events, thoughts about justice and the law naturally arise. “Will he be sued?” “Will she be compensated?” These thoughts might influence how people think about the harm in question. Intent is often a

key determinant of monetary compensation for harms. Therefore, intent can change the monetary value of a harm. Of course, this does not provide a “rational” explanation for what we have characterized as a biased perception of harm (since participants overestimated objectively determined quantities). Moreover, this suggestion is far from mutually exclusive with a motivational account of these findings. Most people probably *want* more compensation to be provided to victims of intentional harms. In fact, it would be difficult to make sense of legal concepts like “punitive damages” if this were not the case.

Blame motivation as the psychological engine of harm magnification

The proposed mechanism of blame motivation implies that differences in harmdoer intent are not *necessary* to produce harm magnification—that the effect is driven by blame motivation, not by intent *per se*. However, the identification of a psychological mechanism does not necessarily make the relationship between intent and harm any less interesting. By way of analogy, classic research on stereotype threat (Steele & Aronson, 1995) shows that making a minority student’s race or gender salient in a classroom can devastate their academic performance. The fact that this effect is mediated, at least in part, by disruptions in working memory processes (Schmader & Johns, 2003) does not make the relationship between race and academic performance any less interesting (or practically important from a policy perspective). It does, however, provide theoretical inroads for interventions and further discovery. We hope the same is true here. To this end, ongoing work seeks to further articulate the psychological mechanisms underlying the magnification of intentional harm (e.g., manipulating of the proposed [psychological] mediator independently of the [structural] difference between intentional and unintentional harm).

REFERENCES

- Ajzen I. & Fishbein, M. (1980) *Understanding attitudes and predicting social behavior*
- Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, 63, 368–378.
- Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126, 556–574.
- Alicke, M. D., & Davis, T. L. (1990). Capacity responsibility in social evaluation. *Personality and Social Psychology Bulletin*, 16, 465–474.
- Ames, D. L. & Fiske, S. T., (2013a). Intentional harms are worse, even when they're not. *Psychological Science*, 24(9), 1755-1762.
- Ames, D. L. & Fiske, S. T., (2013b). Outcome dependency alters the neural substrates of impression formation. *Neuroimage*, 83, 599–608.
- Ames, D. L., Fiske, S. T., & Todorov, A. (2011). Impression formation: A focus on others' intents. In J. Decety & J. Cacioppo (Eds.), *The Handbook of Social Neuroscience*. Oxford University Press, 419-433.
- Ames, D. L., Jenkins, A. J., Banaji, M. R., & Mitchell, J. P. (2008). Taking another's perspective increases self-referential neural processing. *Psychological Science*, 19, 642-644.
- Ames, D. R. (2004). Inside the Mind-Reader's Toolkit: Projection and Stereotyping in Mental State Inference. *Journal of Personality and Social Psychology* 87(3), 340-53.
- Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7(4), 268–277.
- Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual

- perception. *Journal of Personality and Social Psychology*, 91(4), 612-625.
- Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14(3), 110–118.
- Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, 29(5), 832–843.
- Baron-Cohen, S. (1991). Do people with autism understand what causes emotion? *Child development*, 62(2), 385–395.
- Baron-Cohen, S., Leslie, A.M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21, 37-46.
- Baron-Cohen, S., Leslie, A. M. & Frith, U. Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology*, 4, 113-125 (1986).
- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3), 351-368.
- Birch, S., A. (2005). When knowledge is a curse: Children's' and adults' reasoning about mental states. *Current Directions in Psychological Science*, 14, 25-29.
- Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, 18(5), 382–386.
- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, 61, 27–41.
- Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1), B25–B31.

- Borgida, E., & Fiske, S. T. (2008). *Beyond common sense: Psychological science in the courtroom*. New York, NY: Wiley-Blackwell.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3–5.
- Bullock J, Green D, Shang H (2010). Yes, but what's the mechanism (Don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550-558.
- Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, 97, 1232–1254.
- Carlsmith, K.M., Darley, J.M., & Robinson, P.H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83, 284–299.
- Centers for Medicare and Medicaid services (2013). New data show antipsychotic drug use is down in nursing homes nationwide [Press release] Retrieved from <http://www.cms.gov>.
- Chiang, J. J., Eisenberger, N. I., Seeman, T. E., & Taylor, S. E. (2012). Negative and competitive social interactions are related to heightened proinflammatory cytokine activity. *Proceedings of the National Academy of Sciences*, 109, 1878-1882.
- Chen Y, Briesacher B. A., Field T. S., Tjia J, Lau DT, Gurwitz JH (2010). Unexplained variation across US nursing homes in antipsychotic prescribing rates. *Archives of Internal Medicine*, 170(1), 89–95.
- Cohen, M. A. (1988). Corporate crime and punishment: A study of social harm and sentencing practice in the Federal Courts, 1984-1987. *American Criminal Law Review*, 26, 605–660.
- Converse, B. A., Lin, S., Keysar, K, & Epley, N. (2008) In the mood to get over yourself: Mood affects theory-of-mind use. *Emotion*, 8(5), 725-730.

Cosmides, L. & Tooby, J. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology*, 64, 201-229.

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, 1, 568-570.

Darley, J. M., & Gromet, D. M. (2010). The psychology of punishment: Intuition and reason, retribution and restoration. In R. Bobocel, A. Kay, M. Zanna, & J. Olson (Eds.) *The psychology of justice and legitimacy: The Ontario Symposium* (Vol. 11). 229-250
Philadelphia: Psychology Press.

Darley, J. M., & Huff, C. W. (1990). Heightened damage assessment as a result of the intentionality of the damage-causing act. *British Journal of Social Psychology*, 29, 181–188.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, 7, 324–336.

DeGrandpre RJ, Bickel WK, Hughes JR, Layng MP, Badger G (1993). Unit price as a useful metric in analyzing effects of reinforcer magnitude. *Journal of Experimental Anal Behavior* 60(3):641–666.

De Vega, M., Diaz, J. M., & León, I. (1997). To know or not to know: Comprehending protagonists' beliefs and their emotional consequences. *Discourse Processes*, 23(2), 169–192.

Diamond, A., & Kirkham, N. (2005). Not quite as grown-up as we like to think parallels between cognition in childhood and adulthood. *Psychological Science*, 16(4), 291–297.

Dijksterhuis, A., van Knippenberg, A., Kruglanski, A.W. & Schaper, C. (1996). Motivated

- social cognition: Need for closure effects on memory and judgments. *Journal of Experimental Social Psychology*, 32, 254-270.
- Ditto, P. H., & Boardman, A. F. (1995). Perceived accuracy of favorable and unfavorable psychological feedback. *Basic and Applied Social Psychology*, 16, 137–157.
- Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In B. H. Ross (Series Ed.) & D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Vol. Eds.), *Psychology of learning and motivation: Vol. 50. Moral judgment and decision making* (pp. 307–338). San Diego, CA: Academic Press.
- Dunbar, R. I. M., 1998. The social brain hypothesis. *Evolutionary Anthropology* 6, 178-190.
- Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Youngblade, L. (1991). Young children's understanding of other people's feelings and beliefs: individual differences and their antecedents. *Child Development*, 62, 1353-1366.
- Dunning, D., Leuenberger, A., & Sherman, D. A. (1995). A new look at motivated inference: Are self-serving theories of success a product of motivational forces? *Journal of Personality and Social Psychology*, 69, 58–68.
- Eisenberger, N. I., Lieberman, M. D., Williams, K. D., (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, 302, 290-292.
- Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, 87, 327–339.
- Epley, N., & Waytz, A. (2010). Mind Perception. In S.T. Fiske, D.T. Gilbert, & G. Lindsay (Eds.), *The Handbook of Social Psychology* (5th ed., pp. 498-541). New York: Wiley
- Dunning, D. (1999). A newer look: Motivated social cognition and the schematic representation of social concepts. *Psychological Inquiry*, 10(1), 1–11.

- Eisenberger, N. I. (2012). The pain of social disconnection: Examining the shared neural underpinnings of physical and social pain. *Nature Reviews Neuroscience*, 13, 421-434.
- Ferster CB, Skinner, BF (1957). *Schedules of reinforcement* (Appleton-Century-Crofts, New York).
- Fiedler K, Schott M, Meiser T (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology* 47(6), 1231–1236.
- Finkel, E. J., Rusbult, C. E., Kumashiro, M., & Hannon, P. A. (2002). Dealing with betrayal in close relationships: Does commitment promote forgiveness? *Journal of Personality and Social Psychology*, 82(6), 956–974.
- Fiske, S. T. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J. Uleman & J. Bargh (Eds.), *Unintended thought: The limits of awareness, intention, and control* (pp. 253–283). New York, NY: Guilford Press.
- Fiske S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, 38, 889-906.
- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2), 77–83.
- Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6), 878.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1–74). San Diego, CA: Academic Press.

- Fiske, S. T., & Taylor, S. E. (2013). *Social Cognition: From Brains to Culture*. London: Sage
- Gebotys, R., & Dasgupta, B. (1987). Attribution of responsibility and crime seriousness. *The Journal of Psychology*, 121, 607–613.
- Gilbert, D. (2011). Buried by bad decisions. *Nature*, 474, 275–277.
- Gilbert D. T., & Malone P. S. (1995). The correspondence bias. *Psychological Bulletin* 117, 21-38.
- Gollwitzer, M., & Rothmund, T. (2009). When the need to trust results in unethical behavior: The Sensitivity to Mean Intentions (SeMI) model. In D. De Cremer (Ed.), *Psychological perspectives on ethical behavior and decision making* (pp. 135–152). Charlotte, NC: Information Age.
- Gray, K., & Wegner, D. M. (2008). The Sting of Intentional Pain. *Psychological Science*, 19(12), 1260-1262.
- Gray, K., & Wegner, D. M. (2010). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review*, 14, 7–16.
- Gray, K., & Wegner, D. M. (2010). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review*, 14, 7–16.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.
- Haidt, J., & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 797–832). Hoboken, NJ: Wiley.
- Harris, P., Johnson, C. N., Hutton, D., Andrews, G., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition and Emotion*, 3, 379-400.
- Hoffman, E., McCabe, K., & Smith, V. (1998). Behavioral foundations of reciprocity:

- Experimental economics and evolutionary psychology. *Economic Inquiry*, 36, 335–352.
- Holmes, O. W. 1881. *The Common Law*. Macmillan. London.
- Hooker, C. I., Verosky, S. C., Germine, L. T., Knight, R. T., & D’Esposito, M. (2008).
Mentalizing about emotion and its relationship to empathy. *Social cognitive and affective neuroscience*, 3(3), 204–217.
- Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3), 399–425.
- Horton, S., Begin, F., Greig, A., & Lakshman A. (2008). *Micronutrient supplements for child survival (vitamin a and zinc)*. Copenhagen Consensus Center: Denmark.
- Hughes, B.L., & Beer, J.S. (2012). Orbitofrontal cortex and anterior cingulate cortex are modulated by motivated social cognition. *Cerebral Cortex*, 22, 1372-1381.
- Hughes, C., & Dunn, J. (1998). Understanding and emotion: Longitudinal associations with mental-state talk between young friends. *Developmental Psychology*, 34, 1028-1037.
- Hursh S. R., Raslear D., Shurtleff R., Bauman R., & Simmons L. (1988). A cost-benefit analysis of demand for food. *Journal of Experimental Animal Behavior* 50:419-440.
- Imai K., Keele L., & Tingley D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309-334.
- Jenkins, A.C. & Mitchell, J.P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience*, 6(3), 211-218
- Jones E. E., Davis K. E., & (1965). From acts to dispositions: The attribution process in person-perception. *Advances in Experimental Social Psychology* 2, 219-266.
- Keller, J. (2005). In genes we trust: the biological component of psychological essentialism and its relationship to mechanisms of motivated social cognition. *Journal of Personality and*

- Social Psychology*, 88(4), 686-702.
- Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89, 25-41.
- Kinderman, P., Dunbar, R.I.M., Bentall, R.P., 1998. Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, 89, 191-204.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63, 190–193.
- Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., & Sommers, T. (2012). Experimental philosophy. *Annual Review of Psychology*, 63, 81–99.
- Kruglanski, A. W. (1996). Motivated social cognition: Principles of the interface. In, E.T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles*. (pp. 493-520). New York, NY, US: Guilford Press.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3), 480-498.
- Kurzban, R., & Neuberg, S. L. (2005). Managing ingroup and outgroup relationships. In D. M. Buss (Ed.), *Handbook of evolutionary psychology* (pp. 653–675). New York: Wiley.
- Lerner, M. (1965). Evaluation of performance as a function of performer's reward and attractiveness. *Journal of Personality and Social Psychology*, 26, 415–419.
- Lattal K. A., & Gleeson, S. (1990). Response acquisition with delayed reinforcement. *Journal of Experimental Psychology: Animal Behavioral Processes*, 16(1), 27-39.
- Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition*, 50, 193-200.
- Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret

- behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46, 551-556.
- Loftus, E. F. (1997). Creating false memories. *Scientific American*, 277, 70–75.
- Malle, B. F. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. In J. M. Olson & M. P. Zanna (Eds.), *Advances of Experimental Social Psychology* (Vol. 44, pp. 297-352). Burlington: Academic Press.
- Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, 102, 661–684.
- Malone C., & Fiske S. T. (2013). *The human brand: How we relate to people, products and companies* (Wiley/Jossey Bass: San Francisco).
- McGraw, A. P., Todorov, A., & Kunreuther, H. (2011). A policy maker's dilemma: Preventing terrorism or preventing blame. *Organizational Behavioral and Human Decision Processes*, 115, 25-34.
- Morewedge C. K., Huh Y. E., & Vosgerau J. (2010). Thought for food: Imagined consumption reduces actual consumption. *Science*, 330(6010), 1530–1533.
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44, 1–23.
- Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin*, 135(5), 749-773.
- Mitchell, J. P. (2009). Inferences about other minds. *Philosophical Transactions of the Royal Society B*, 364, 1309-1316.

- Murray, M. (2007). Narrative Psychology. In J. A. Smith (Ed.) *Qualitative Psychology: A Practical Guide to Research Methods*, pp. 111-131. Sage.
- Newton, L. (1990). "Overconfidence in the Communication of Intent: Heard and Unheard Melodies," Ph.D. dissertation.
- Newton, A. M., & de Villiers, J. G. (2007). Thinking While Talking Adults Fail Nonverbal False-Belief Reasoning. *Psychological Science*, 18(7), 574–579.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872.
- Paolacci, G., Chandler, J., & Ipiertis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419.
- Pe'er, E., Paolacci, G., Chandler, J., & Mueller, P., (2012) Selectively Recruiting Participants from Amazon Mechanical Turk Using Qualtrics. Available at SSRN: <http://ssrn.com/abstract=2100631>.
- Perner, J., Ruffman, T. & Leekam, S. R. (1994). Theory of mind is contagious: you catch it from your sibs. *Child Development*, 65, 1228-1238.
- Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3), 437–471.
- Pizarro, D. A., Laney, C., Morris, E. K., & Loftus, E. F. (2006). Ripple effects in memory: Judgments of moral blame can distort memory for events. *Memory & Cognition*, 34, 550–555.

- Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 20, pp. 297–340). New York, NY: Academic Press.
- Reeder, G. D., & Brewer M.B. (1979) A schematic model of dispositional attribution in interpersonal perception. *Psychological Review* 86, 61-79.
- Robinson, P. H., & Darley, J. M. (1995). *Justice, liability, and blame: Community views and the criminal law*. Boulder, CO: Westview Press.
- Roets, A., Van Hiel, A., & Dhont, K. (2012). Is Sexism a Gender Issue? A Motivated Social Cognition Perspective on Men's and Women's Sexist Attitudes Toward Own and Other Gender. *European Journal of Personality*, 26(3), 350–359.
- Rothmund, T., Gollwitzer, M., & Klimmt, C. (2011). Of virtual victims and victimized virtues: Differential effects of experienced aggression in video games on social cooperation. *Personality and Social Psychology Bulletin*, 37, 107–119.
- Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, 5/6, 359-371.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19, 1835-1842.
- Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, 85(3), 440-452.

- Shaver, K. G. (1970). Defensive attribution: Effects of severity and relevance on the responsibility assigned for an accident. *Journal of Personality and Social Psychology*, 14, 101–113.
- Showers, C., & Cantor, N. (1985). Social cognition: A look at motivated strategies. *Annual review of psychology*, 36(1), 275–305.
- Sevillano V., & Fiske, S. T. (2013). Perceived dimensions of animals. Poster presented at the Annual meeting of the Society for Personality and Social Psychology, 17-19 January. New Orleans, LA.
- Sodian, B. (2011). Theory of Mind in Infancy. *Child Development Perspectives*, 5(1), 39–43.
- Span, P. (2010) Overmedication in the nursing home. *The New York Times* Retrieved from <http://www.nytimes.com>.
- Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology* 69(5), 797–811.
- Stevens, L. E., & Fiske, S. T. (1995). Motivation and cognition in social life: A social survival perspective. *Social Cognition*, 13(3), 189–214.
- Stiller, J., & Dunbar, R. I. M., 2007. Perspective-taking and social network size in humans. *Social Networks*, 29, 93-104.
- Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, 30, 395– 402.
- United States Congressional Budget Office (2010). CBO's Analysis of Scenarios for Funding the Wars in Afghanistan and Iraq. Washington. <http://www.cbo.gov/publication/42001>
- USAID Micronutrient Program. 2004. *Cost Analysis of the National Vitamin A Supplementation*

Programs in Ghana, Nepal, and Zambia: A Synthesis of Three Studies. Arlington, Virginia, USA

- Vangelisti, A. L., & Young, S. L. (2000). When Words Hurt: The Effects of Perceived Intentionality on Interpersonal Relationships. *Journal of Social and Personal Relationships, 17*(3), 393–424.
- Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829-858.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child development, 72*(3), 655–684.
- Willis, A., & Todorov, (2006). *First impressions: Making up your mind after 100-ms exposure to a face. Psychological Science. 17, 592-589.*
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.
- Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21*, 1396–1405.
- Youngblade, L., M., & Dunn, J. (1995). Individual Differences in young children's pretend play with mother and sibling: Links to relationships and understanding of other people's feelings and beliefs. *Child Development, 66*, 1472-1492.
- Zaki, J. & Ochsner, K. (2012). The neuroscience of empathy: Progress, pitfalls, and promise. *Nature Neuroscience, 15*(5), 675-680.

Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy.
Proceedings of the National Academy of Sciences, 106(27), 11382–11387.

Appendix A: Participant Exclusion Criteria – Chapter 1

For Studies 1-3, all 294 participants were included in all analyses. Studies 4 and 5 included basic data quality checks. Participants who failed any one of the following criteria tended to fail others as well; thus the *ns* reported below overlap heavily. Participants were excluded from Studies 4-5 if they (a) did not complete the task, i.e., provided no estimate ($n=39$), (b), provided an estimate too small to be logically valid (this was operationalized as an estimate for the sum of the seven numbers that was not greater than the largest one of those seven numbers, and therefore, not a logically possible sum; ($n=49$), (c) provided an estimate so large that we did not believe it represented a genuine estimate (four such estimates were excluded: \$80,000 \$100,000, \$225,000 and \$108.2 million), (d) spent fewer than five seconds “reading” the instructions for the main task ($n=39$), (e) failed the manipulation check (“Was the river drying up caused by a person?” [yes/no]) ($n=44$), (f) failed the reading check question (“For how many weeks did the river dry up?”) (correct answer, 2 weeks—participants were considered to have failed if they were off by more than one week, though results were statistically identical if this criterion was made more stringent so as to require a precisely correct answer, or if the criterion was relaxed or removed altogether; $n=19$). Following the application of these a priori criteria, participants were also excluded as statistical outliers if their estimates were more than three standard deviations above or below the mean ($n=7$).

In addition, participants who had participated in any study reported here or in any study related to this line of research were automatically and preemptively disqualified from all other studies related to this line of research. This was accomplished via newly developed and publicly available code (Pe’er, Paolacci, Chandler & Mueller, 2012).

Reference

Peer, E., Paolacci, G., Chandler, J., & Mueller, P. (2012). Screening participants from previous studies on Amazon Mechanical Turk and Qualtrics. Retrieved June 27, 2012 from <http://experimentalturk.wordpress.com/resources/>

Appendix B: Complete Vignette Text – Chapter 1

Studies 1-3, Unintentional-Harm Condition:

Terrance Smith is the CEO of a medium-sized company. Part of his employees' pay comes from sharing company profits. This system carries some risk for the employees, but they like it, and it motivates them to work hard. Partly because of this motivation, the company is starting to become exceptionally profitable.

Recently, Terrance decided to invest in a software development project because it seemed like a good opportunity for the company. Unfortunately, not all investments are successful.

The investment failed. The employees continued to do their jobs as always without paying attention to the details of the company's investments. Neither they, nor anyone else (besides Terrance) knew why profits were lower than expected. Terrance's own salary was unaffected, but the employees made less money than they otherwise would have. This caused them some anxiety and made paying off credit cards and taking family vacations harder for them. Two employees saw a minor dip in their credit scores because of late payments— though nobody suffered truly dire financial hardship or came close to losing their home or car or anything. Terrance was largely unaware of these specific consequences in his employees' personal lives, since he didn't talk to them about personal matters very much.

Studies 1-3, Intentional-Harm Condition

Terrance Smith is the CEO of a medium-sized company. Part of his employees' pay comes from sharing company profits. This system carries some risk for the employees, but they like it, and it motivates them to work hard. Partly because of this motivation, the company is starting to become exceptionally profitable.

Recently, Terrance decided to invest in a software development project because he was confident that it would fail. He thought his employees might become lazy now that the company was starting to do well, and that they would work harder to increase profits in the future if they made less money now.

The investment failed. The employees continued to do their jobs as always without paying attention to the details of the company's investments. Neither they, nor anyone else (besides Terrance) knew why profits were lower than expected. Terrance's own salary was unaffected, but the employees made less money than they otherwise would have. This caused them some anxiety and made paying off credit cards and taking family vacations harder for them. Two employees saw a minor dip in their credit scores because of late payments— though nobody suffered truly dire financial hardship or came close to losing their home or car or anything. Terrance was largely unaware of these specific consequences in his employees' personal lives, since he didn't talk to them about personal matters very much.

Study 3, Intentional-Harm/Caught Condition (added to end of intentional condition vignette)

Soon after, Terrance was caught. He was immediately fired and had to pay a fine (a large one, even for a CEO) to the Consumer Financial Protection Bureau. He was also publicly blamed and morally condemned in the media, which negatively impacted his personal and professional life.

Studies 4 & 5, Unintentional-Harm Condition

There is a small group of 20 people living in the desert. They are mostly self-sufficient, growing their own food using simple irrigation techniques. For this reason, their wellbeing is

dependent on the river flowing through their encampment.

The river dried up for a period of two weeks due to a lack of rain in the mountains.

The people did not know why their river dried up, but it had a number of consequences. Although they always kept some water in reserve, they didn't have enough for two weeks, and about 20% of their crops failed. Everyone was thirsty and miserable, three people got seriously sick from dehydration, though nobody actually died. After two weeks, the river began flowing normally again. Though everyone suffered, some members of the community suggested that the experience had a silver lining because it tested their resilience, and brought them closer together.

Studies 4 & 5, Intentional-Harm Condition

There is a small group of 20 people living in the desert. They are mostly self-sufficient, growing their own food using simple irrigation techniques. For this reason, their wellbeing is dependent on the river flowing through their encampment.

The river dried up for a period of two weeks because a man who lived in a nearby town diverted the flow. He did this because it made him feel powerful and he thought it would be fun to observe them struggling using his high-powered binoculars.

The people did not know why their river dried up, but it had a number of consequences. Although they always kept some water in reserve, they didn't have enough for two weeks, and about 20% of their crops failed. Everyone was thirsty and miserable, three people got seriously sick from dehydration, though nobody actually died. After two weeks, the river began flowing normally again. Though everyone suffered, some members of the community suggested that the experience had a silver lining because it tested their resilience, and brought them closer together.

Appendix C: Supplemental mediation analyses – Chapter 1

Study 1

Nonparametric bootstrapping is arguably a superior approach to establishing mediation as compared with the Sobel approach (see Preacher & Hayes, 2004; Preacher, Rucker, & Hayes, 2007). We therefore tested the predicted model of blame motivation as a mediator of the relationship between perceived intent and perceived harm based on 10,000 bootstrapped samples. Mediation is considered to be significant if the 95% bias corrected and accelerated confidence intervals for the indirect effect do not include zero (Preacher & Hayes, 2004; Preacher et al., 2007). Results fully corroborated the Sobel results, as reported in the main text. The total effect of intent on harm was significant ($TE = 0.25$, $SE = 0.07$, $p < .001$), while the direct effect was not ($DE = -.14$, $SE = 0.09$, $p = .13$). Blame motivation thus fully mediated the relationship between intent and harm (IE lower 95% CI=.43, upper 95% CI=.78, fully standardized), such that participants who saw the CEO's actions as highly intentional were more motivated to blame, and through this blame motivation, saw more harm. Because zero is not in the 95% confidence interval, the indirect effect is significantly different from zero at $p < .05$ (two tailed).

As noted in the main text, both the Sobel method and the bootstrapping method support the predicted model (intent→blame motivation→harm) over the alternative model (intent→harm→blame motivation). In both cases, the predicted model yields full mediation, while the alternative model yields relatively weak and partial mediation. In a “traditional” analysis of the alternative model, the residual direct effect of intent on blame motivation remained significant (and was, in fact, the strongest path in the model) after accounting for perceived harm as a possible mediator, standardized $\beta = .61$, $p < .001$. Likewise, based on 10,000 bootstrapped samples of the alternative model, the direct effect of intent on blame

motivation remained significant after accounting for perceived harm as a potential mediator ($DE = .47, SE = .05, p < .001$).

Study 3

Mediation results for Study 3 were essentially identical to those observed in Study 1. Bootstrapping results again aligned with Sobel results, and again argued for the predicted model over the alternative model. Under the predicted model, the total effect of intent on harm was significant ($TE = 0.17, SE = 0.05, p < .001$), while the direct effect was not ($DE = .08, SE = 0.05, p = .14$). Thus, blame motivation fully mediated the relationship between intent and harm (IE lower 95% $CI = .08$, upper 95% $CI = .30$, fully standardized).

Also as in Study 1, both mediation approaches support the predicted model (intent → blame motivation → harm) over the alternative model (intent → harm → blame motivation). In both cases, the predicted model yields full mediation, while the alternative model yields only partial mediation. In a “traditional” analysis, the residual direct effect of intent on blame motivation remained significant (and was, again, the strongest path in the model) after accounting for perceived harm as a potential mediator, standardized $\beta = .39, p < .001$. Likewise, based on 10,000 bootstrapped samples of the alternative model, the direct effect of intent on blame motivation remained significant after accounting for perceived harm as a potential mediator ($DE = .30, SE = .06, p < .001$). Results were essentially identical if participants from the intentional-caught condition were excluded from the analysis.

References

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*,

36, 717-731.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: *Theory, methods, and prescriptions*. *Multivariate Behavioral Research*, 42(1), 185–227.

Appendix D: Complete Vignette Text – Chapter 2

Study 1: Modified Sally-Anne Task

Sally and Anne were washing dishes together. Sally took off her ring and put it in the soap dish. The ring was cheap and it didn't have special sentimental value to her or anything--but sometimes she wore it anyway.

Sally and Anne then went back to washing dishes and talking. A minute later, Sally's phone rang. It was her dad. Sally apologized to Anne and asked if it was okay for her to go talk to her dad for a few minutes. Anne said that it was totally fine, and that she (Anne) didn't mind finishing the rest of the dishes without Sally, and that she would just leave when she was done.

Once Sally left, Anne noticed that Sally's ring was still in the soap dish. Sally tended not to pay attention to where she put things, and it seemed like the ring could fall into the sink pretty easily. Anne thought this would make Sally upset.

Anne [**wanted/didn't want**] her friend Sally to be upset. She actually didn't like Sally all that much sometimes, [**and/but**] she [**wanted/didn't want**] to make her to feel bad. Anne decided to move Sally's cheap ring to a drawer so that [**Sally would think that it had gone/it wouldn't go**] down the drain.

Anne then went back to the sink, finished washing and drying the dishes, and went home.

Appendix E: Sally-Anne Graphics – Chapter 2

[Participants saw the same set of illustrations in the same order across the two conditions.]



A. Sally and Anne washing dishes together



B. Sally putting down her ring



C. Sally and Anne returning to washing dishes



D. Sally's phone ringing



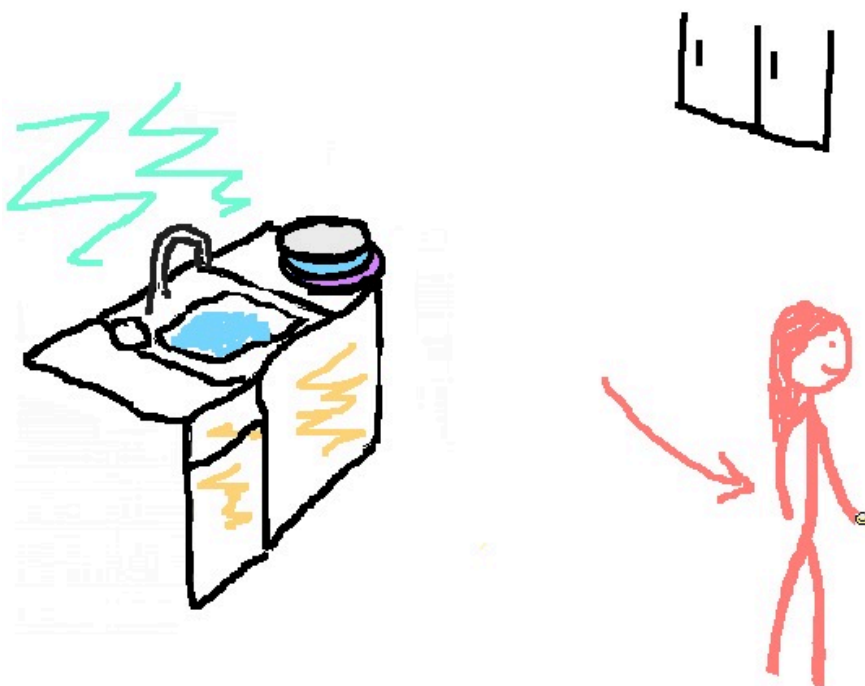
E. Sally leaving the room



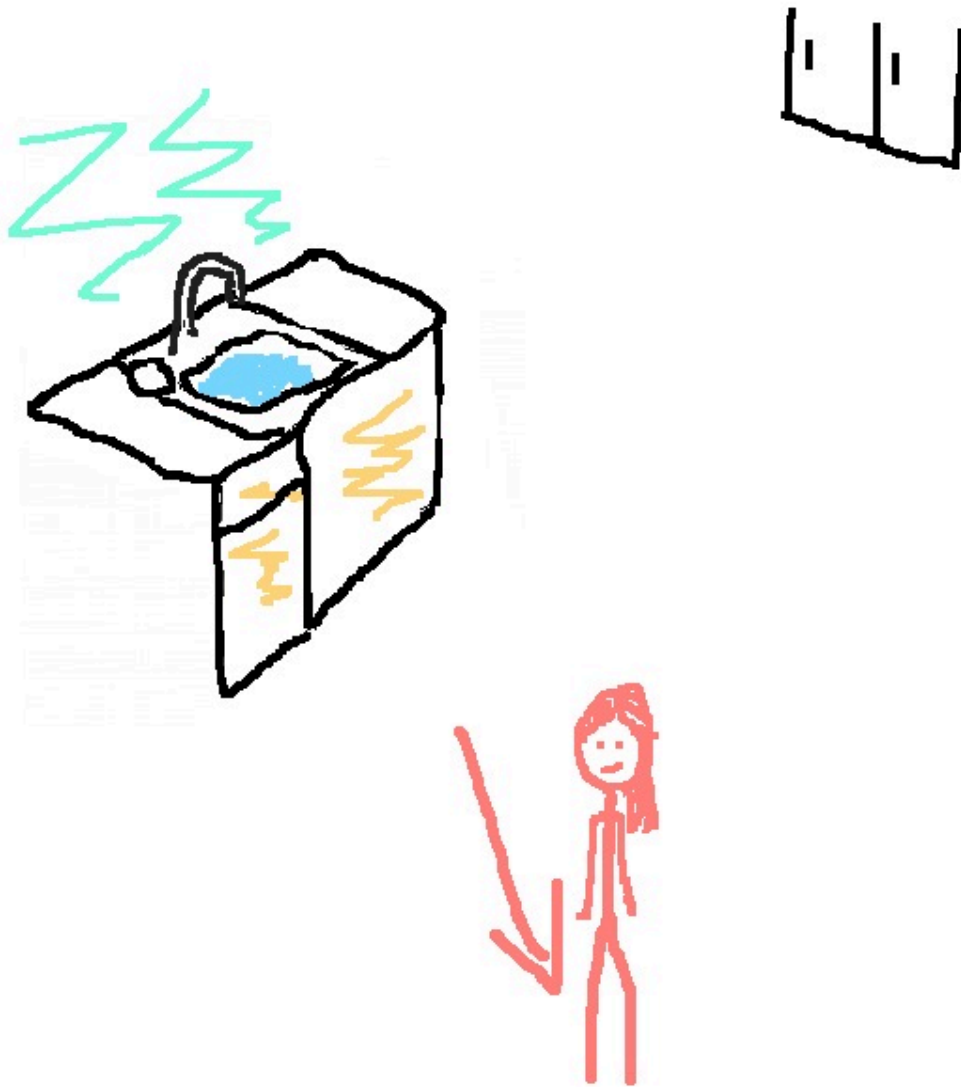
F. Anne noticing Sally's ring



G. Anne picking up Sally's ring



H. Anne moving Sally's ring



I. Anne leaving

Appendix F: Supplementary Mediation Analyses – Chapter 2

Study 1 – Modified Sally-Anne task:

Bootstrapping procedures (1,000 samples) were employed to both the proposed mediation model (intent→blame motivation→harm) and the alternative model (intent→harm→blame motivation). In the proposed model, intent (IV) was positively associated with perceived harm (DV), (*c*-path; unstandardized $\beta = .10 (.05)$, $t(114) = 2.16$, $p = .03$). Intent was also positively related to the mediator, blame motivation (*a*-path; unstandardized $\beta = .68 (.03)$, $t(114) = 22.44$, $p < .001$). In addition, the mediator, blame motivation, was positively associated with the dependent variable, perceived harm (*b*-path; unstandardized $\beta = .39 (.14)$, $t(114) = 2.91$, $p = .004$). Having verified that the *a*, *b*, and *c*-paths were all significant, we tested for mediation using bootstrapping methods with bias-corrected confidence estimates (Preacher & Hayes, 2004; Preacher, Rucker & Hayes, 2007). This analysis confirmed the mediating role of blame motivation in the effect of intent on perceived harm (unstandardized β of indirect effect = $.27 (.09)$, 95% CI $[.090, .457]$; completely standardized β of indirect effect = $.54 (.18)$, 95% CI $[.176, .889]$). Because zero is not in the 95% confidence interval, the indirect effect is considered significantly different from zero at $p < .05$ (two tailed). In addition, the residual relationship between the IV and DV after controlling for the mediator was nonsignificant. This is indicated by examining the direct effect of the IV, intent, on the DV, perceived harm, while controlling for the mediator, blame motivation (*c'*-path; unstandardized $\beta = -.17$, $t(114) = 1.67$, $p = .10$). As noted in the main text, this is conventionally taken as evidence of “full mediation;” but the traditional distinction between “full” and “partial” mediation is likely a tenuous one (Rucker et al., 2011).

By comparison, the alternative mediation model (intent→harm→blame motivation) performed poorly. First, the Sobel test for the alternative model was nonsignificant, $Z = 1.67$ (.01), $p = .09$, and the coefficients for the total and direct effects in this model (c and c' paths) were virtually identical (respectively .68 vs. .67, unstandardized, $ps < .001$), whereas the c' path dropped to nonsignificance ($p = .10$) in the predicted model. This suggests that, while blame motivation as a mediator accounted reasonably well for the relationship between intent and perceived harm (predicted model), perceived harm did not account for much, if any, of the statistical relationship between intent and blame motivation (alternative model). This inference is confirmed by examination of the indirect effect for the alternative model ($\beta = .02$ (.01), CI [.003, .046] unstandardized; $\beta = .02$ (.01), 95% CI [.004, .062], completely standardized).

Study 3: Vlog

Mediation results for Study 3 were essentially identical to those observed in Study 1. In the proposed model, intent (IV) was positively associated with perceived harm (DV), (c -path; unstandardized $\beta = .12$ (.05), $t(101) = 2.59$, $p = .01$). Intent was also positively related to the mediator, blame motivation (a -path; unstandardized $\beta = .65$ (.05), $t(101) = 14.15$, $p < .001$). In addition, the mediator, blame motivation, was positively associated with the dependent variable, perceived harm (b -path; unstandardized $\beta = .32$ (.10), $t(101) = 3.22$, $p = .002$). Having verified that the a , b , and c -paths were all significant, we tested for mediation using bootstrapping methods with bias-corrected confidence estimates (Preacher & Hayes, 2004; Preacher, Rucker & Hayes, 2007). This analysis confirmed the mediating role of blame motivation in the effect of intent on perceived harm (unstandardized β of indirect effect = .21 (.06), 95% CI [.089, .350], 1,000 samples; completely standardized β of indirect effect = .42 (.11), 95% CI [.190, .651].

Because zero is not in the 95% confidence interval, the indirect effect is considered significantly different from zero at $p < .05$ (two tailed). In addition, the residual relationship between the IV and DV after controlling for the mediator was nonsignificant. This is indicated by examining the direct effect of the IV, intent, on the DV, perceived harm, while controlling for the mediator, blame motivation (c' -path; unstandardized $\beta = -.08$, $t(101) = 1.05$, $p = .30$).

By comparison, the alternative mediation model (intent $\rightarrow$ harm $\rightarrow$ blame motivation) performed poorly. First, the Sobel test for the alternative model was only marginally significant, $Z = 1.96$ (.02), $p = .05$, and the coefficients for the total and direct effects in this model (c and c' paths) were virtually identical (respectively .65 vs. .62, unstandardized, $ps < .001$), whereas the c' path dropped to nonsignificance in the predicted model. This suggests that, while blame motivation as a mediator accounted reasonably well for the relationship between intent and perceived harm (predicted model), perceived harm did not account for much, if any, of the statistical relationship between intent and blame motivation (alternative model). This inference is confirmed by examination of the indirect effect for the alternative model ($\beta = .04$ (.02), CI [.004, .097] unstandardized; $\beta = .05$ (.03), 95% CI [.005, .115], completely standardized).

Study 3: Vlog replication:

Mediation results for the Study 3 replication were essentially identical to those observed using the original materials. In the proposed model, intent (IV) was positively associated with perceived harm (DV), (c -path; unstandardized $\beta = .18$ (.04), $t(198) = 4.72$, $p < .001$). Intent was also positively related to the mediator, blame motivation (a -path; unstandardized $\beta = .63$ (.03), $t(198) = 20.9$, $p < .001$). In addition, the mediator, blame motivation, was positively associated with the dependent variable, perceived harm (b -path; unstandardized $\beta = .30$ (.05), $t(198) = 5.90$,

$p < .001$). Having verified that the a , b , and c -paths were all significant, we tested for mediation using bootstrapping methods with bias-corrected confidence estimates (Preacher & Hayes, 2004; Preacher, Rucker & Hayes, 2007). This analysis confirmed the mediating role of blame motivation in the effect of intent on perceived harm (unstandardized β of indirect effect = .19 (.06), 95% CI [.086, .322]; completely standardized β of indirect effect = .33 (.10), 95% CI [.156, .539]. Because zero is not in the 95% confidence interval, the indirect effect is considered significantly different from zero at $p < .05$ (two tailed). In addition, the residual relationship between the IV and DV after controlling for the mediator was nonsignificant. This is indicated by examining the direct effect of the IV, intent, on the DV, perceived harm, while controlling for the mediator, blame motivation (c' -path; unstandardized $\beta = -.01$, $t(197) = 0.18$, $p = .85$).

By comparison, the alternative mediation model (intent $\rightarrow$ harm $\rightarrow$ blame motivation) performed less well. First, the coefficients for the total and direct effects in this model (c and c' paths) were virtually identical (respectively .63 vs. .60, unstandardized, $ps < .001$), whereas the c' path dropped to nonsignificance in the predicted model. This suggests that, while blame motivation as a mediator accounted reasonably well for the relationship between intent and perceived harm (predicted model), perceived harm did not account for much of the statistical relationship between intent and blame motivation (alternative model). This inference is confirmed by examination of the indirect effect for the alternative model ($\beta = .04$ (.01), CI [.013, .066] unstandardized; $\beta = .05$ (.02), 95% CI [.017, .087], completely standardized. The Sobel test for the alternative model was significant, $Z = 2.74$ (.02), $p = .006$, though the effect was small (.04).

REFERENCES

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, 36, 717-731.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1), 185–227.

Appendix G: Participant Exclusion Criteria – Chapter 3

Participants and Recruitment

For Studies 2 and 3, all participants were included in all analyses. Study 1 included basic data quality checks. Participants who failed any one of the following criteria sometimes failed others as well; thus the *ns* reported below overlap. Participants were excluded from Study 1 if they (a) did not complete the task, i.e., provided no response to the estimation task, and/or the question concerning whether the act was intentional, and/or the basic reading check question ($N = 27$); (b) provided an estimate too small to be logically valid (this was operationalized as an estimate for the sum of the eight numbers that was not greater than the largest one of those eight numbers, and therefore, not a logically possible sum ($N = 10$); (c) provided an estimate so large that we did not believe it represented a genuine estimate (one such estimate was excluded: \$150,000—two orders of magnitude above the correct answer of \$4,433.43); (d) failed the manipulation check (i.e., were unable to report correctly whether the act was intentional or unintentional) ($N = 8$); (e) failed the reading check question (Did all of the residents recover in the end?"; correct answer = "yes") ($N = 9$). Following the application of these a priori criteria, participants were also excluded as statistical outliers if their estimates were more than three standard deviations above or below the condition mean ($N = 6$).

Appendix H: Complete Vignette Text – Chapter 3

Intentional harm condition:

Jake F. was a caretaker at the Kind Shepherd Nursing Home in Pennsylvania. One day in early 2013, Jake switched a few of the residents' medications. He did this intentionally because he thought that it would be harmless fun to watch some of the old people be in pain, or not know where they were for a couple of days, etc. Everyone agreed that the residents were sometimes frustrating to deal with, and Jake wanted to hurt some of them a bit.

The specific pills that were switched looked identical, and the manufacturer labeled them poorly; so none of Jake's coworkers had any way to notice the switch. Also, Jake had never mixed up anyone's medications before.

The consequences for the residents were significant, but non-lethal. One resident experienced heart palpitations, became very confused, and spent almost a full day tearfully insisting that she be allowed to see her (deceased) husband. Another had an anaphylactic (severe allergic) reaction, which included a dangerous drop in blood pressure. Several residents needed extra medical treatment which was, at least for the time being, billed to the residents themselves (per the normal contract agreement). In the end, everyone recovered. The residents who were affected did not know why they got sick, but were grateful for the care they received from the staff during recovery.

Unintentional harm condition:

Jake F. was a caretaker at the Kind Shepherd Nursing Home in Pennsylvania. One day in early 2013, Jake accidentally mixed up a few of the residents' medications. He did this completely unintentionally, and not because he thought that it would be harmless fun to watch

some of the old people be in pain, or not know where they were for a couple of days, etc.

Everyone agreed that the residents were sometimes frustrating to deal with, but Jake never wanted to hurt any of them.

The specific pills that were switched looked identical, and the manufacturer labeled them poorly; so none of Jake's coworkers had any way to notice the mix-up. Also, Jake had never mixed up anyone's medications before.

The consequences for the residents were significant, but non-lethal. One resident experienced heart palpitations, became very confused, and spent almost a full day tearfully insisting that she be allowed to see her (deceased) husband. Another had an anaphylactic (severe allergic) reaction, which included a dangerous drop in blood pressure. Several residents needed extra medical treatment which was, at least for the time being, billed to the residents themselves (per the normal contract agreement). In the end, everyone recovered; and the residents who were affected did not know why they got sick, but were grateful for the care they received from the staff during recovery.

Appendix I: Analytical Approach for Clicking Behavior – Chapter 3

Analyses of time spent clicking and click rate include only participants who clicked at least once (0 clicks is valid data for the count-based analysis, but 0 clicks in 30 seconds is uninformative for click rate, and including such data would bias results in a direction consistent with hypotheses). Time spent clicking and click rate were indexed by the times at which participants began and ended the task. This approach underestimates true click rate, because participants presumably spent part of this time reading instructions and orienting to the task. In principle, it is possible that participants in one condition took longer to begin clicking, or remained in the task for longer after they were done clicking. However, this cannot explain the critical results observed in this experiment. If participants in the intentional harm condition spent longer in the task, without clicking more, this could indeed produce an artificial “time spent clicking” difference, but could not explain the “number of clicks” difference (or why it was maintained across every level of reinforcement), or the “click rate” difference (it would predict the opposite of the observed results). Likewise, if participants in the unintentional harm condition spent more time on the task without actually clicking any less, their click rate would, in fact, be artificially deflated in our analysis; however, this cannot explain differences in the number of observed clicks, and suggests that our observed effect of time spent working is, in fact, even larger than what we report.