Optimality Bias in Moral Judgment

Julian De Freitas [1*]
Samuel G. B. Johnson [2*]

[1] Department of Psychology, Harvard University
[2] Division of Marketing, Business, & Society, University of Bath School of Management

*Equal Contributions

Corresponding Author: Julian De Freitas, defreitas@g.harvard.edu

Last Revised: 24 July 2018 (accepted version)

Note. This paper is a pre-print of an article to appear in the *Journal of Experimental Social Psychology*.

Abstract

We often make decisions with incomplete knowledge of their consequences. Might people nonetheless expect others to make optimal choices, despite this ignorance? Here, we show that people are sensitive to *moral optimality*: that people hold moral agents accountable depending on whether they make optimal choices, even when there is no way that the agent could know *which* choice was optimal. This result held up whether the outcome was positive, negative, inevitable, or unknown, and across within-subjects and between-subjects designs. Participants consistently distinguished between optimal and suboptimal choices, but not between suboptimal choices of varying quality — a signature pattern of the Efficiency Principle found in other areas of cognition. A mediation analysis revealed that the optimality effect occurs because people find suboptimal choices more difficult to explain and assign harsher blame accordingly, while moderation analyses found that the effect does not depend on tacit inferences about the agent's knowledge or negligence. We argue that this moral optimality bias operates largely out of awareness, reflects broader tendencies in how humans understand one another's behavior, and has real-world implications.

Keywords: Moral judgment; Lay decision theory; Theory of mind; Decision-making; Causal attribution.

Introduction

We hold others accountable for their actions based on what they were thinking. If a student cheats on an exam, a scientist fabricates a result, or a company mistreats a customer, our judgment depends on their motives and beliefs. Empirical studies confirm this intuition (e.g., Cushman et al., 2013; Gray & Wegner, 2008), and all theories of blame must account for it (e.g., Cushman, 2008; Malle, Guglielmo, & Monroe, 2014; Shaver, 1985; Uhlmann, Pizarro, & Diermeier, 2015). Yet, people often seem to blame others for things they could not possibly have known about. In 2009, a group of seismologists issued a statement indicating that an earthquake in L'Aquila, Italy was unlikely; when an earthquake struck and killed 308 people, they were charged with manslaughter. Despite the defense's insistence that it is simply beyond the powers of science to predict earthquakes, the scientists were sentenced to prison. Although the convictions were ultimately overturned, incidents like this highlight ways in which our moral judgments can sometimes directly contradict inferences about agents' intentions. It is perfectly clear that the scientists did not — and *could not* — know that the earthquake would hit, yet many people blamed the scientists all the same. What psychological principles could explain such paradoxical judgments?

One likely factor is the outcome bias (e.g., Baron & Hershey, 1988) wherein people blame agents for negative *consequences* despite positive intentions. For example, the scientists might have been blamed so long as the earthquake occurred, even if the scientists took pains to avoid making the incorrect prediction that the earthquake would not occur. In real world cases, however, multiple factors are often at play. Not only did the scientists' choice result in a bad *outcome*, but, unknown to the scientists, it was also *suboptimal*. That is, even before the earthquake itself, an omniscient scientist could have known that the earthquake was likely to occur. Thus, the optimal choice would objectively have been to recommend evacuation. Given that scientists and other humans are not omniscient, the scientists did not and could not have *known* that their choice was suboptimal. Yet might people nonetheless blame agents for making suboptimal choices, even when agents have no way of knowing that their choices are suboptimal?

The Efficiency Principle. We propose that moral judgments are influenced by a principle people use for understanding others' behavior, which can override inferences about mental states: People expect agents to behave *optimally* or e*fficiently*, relative to the agent's goals and the constraints of the situation (Dennett, 1987; Gergely & Csibra, 2003). To use an analogy outside of moral judgment, if the car to our right changes lanes, we could understand that decision in terms of the assumed beliefs and desires of the car's driver; but in most cases, we probably use the simpler strategy of understanding the car's behavior in terms of more general features of the world, such as common goals (avoiding collisions) and broad situational constraints (intuitive physics and geometry), and assuming optimal decision-making relative to those constraints. This *Efficiency Principle* runs psychologically deep. It develops before a representational theory of mind (Csibra et al., 1999; Gergely et al., 2003) and may scaffold later-emerging mental-state inferences. It also plays important roles — often outside of awareness — in other domains of cognition, including visual perception (Gao & Scholl, 2011) and language understanding (Davidson, 1967; Grice, 1989).

Most of the time, efficiency-based thinking leads to the same conclusions as mentalizing — after all, people behave in a reasonably rational manner much of the time. For example, imagine that Jill is deciding which of three shampoos to buy, wanting to make her hair smell like apples. Suppose that the three brands have different likelihoods of achieving this goal — one has a 70% efficacy (call this "Best"), one a 50% efficacy ("Middle"), and one a 30% efficacy ("Worst") — and that Jill knows these probabilities. If we think about Jill's mental states, we realize she is most likely to choose Best (since Jill believes, correctly, that this choice is optimal), but we can also reach this conclusion by merely considering what is optimal *in the world*, since Jill's mental states track the world. That is, when an agent's beliefs match the world, efficiency-based thinking is a useful shortcut for predicting behavior. This is why most game theory models assume optimal decision-making from one's opponents (e.g., Morgenstern & von Neumann, 1947; Nash, 1951).

However, there are some situations where normative prediction requires us to override the Efficiency Principle — cases in which the agent is *ignorant* of key information. For example, imagine that Jill is in the same situation as before, but falsely believes that all three shampoos are equally likely to achieve her goal. In this case, our representational theory-of-mind tells us that Jill is equally likely to choose each of the three brands, since she has no reason to choose one over the others. Yet, the Efficiency Principle says that Jill would behave optimally relative to the *true* situational constraints, not relative to her *representation* of those constraints — she would be likely to choose the 70% option, and unlikely to choose the other two options.

Surprisingly, even adults are susceptible to such efficiency-based thinking, which can override theory-of-mind. People believe that Jill, even when ignorant about the relevant probabilities, is most likely to choose the optimal (70%) option, and less likely to choose the suboptimal (50% or 30%) options (Johnson & Rips, 2014). Critically, people also believe that Jill is *equally likely* to choose each of the suboptimal (50% and 30%) options; hence, their predictions track optimality as such, rather than the objective probability of success. This stands in contrast both to normative mental-state inferences (i.e., Jill is equally likely to choose each option) and to the predictions people make for agents who do know the probabilities (i.e., she is more likely to choose Best than Middle, but also more likely to choose Middle than Worst). Thus, this *stepwise pattern* of responses — higher predictions for optimal choices, but roughly equal predictions among different suboptimal choices — is a unique signature of efficiency-based reasoning about ignorant agents. This pattern has been found in both predictions of behavior as well as explanations: People believe that suboptimal choices are more in need of explanation than optimal choices because such choices violate our expectations about optimal behavior, eluding the efficiency-based schema we can typically apply (Johnson & Rips, 2014).

Optimality and morality. These findings led us to predict that suboptimal actions would also lead to (non-normatively) harsher *moral* judgments, in light of people's belief that suboptimal choices are more in need of explanation (Johnson & Rips, 2014). This hypothesis follows from several streams of research.

First, people feel muted affect toward events that are explained (Wilson & Gilbert, 2008) — that is, if they can (intrapersonally) assign meaning to that event. In one study, students studying in the library unexpectedly received a dollar coin attached to an index card. The students maintained a positive mood for a shorter duration when the index card contained text explaining why they had received the gift, compared to when the text on the card eluded explanation (Wilson, Centerbar, Kermer, & Gilbert, 2005). This logic applies to negative events too. Participants encouraged to focus on "why" rather than "what" when recalling an angering experience were less likely to experience negative affect (Kross, Ayduk, & Mischel, 2005). For this reason, people faced with bereavement can cope better with their loss if they are able to find meaning in the death of their loved one (e.g., Bonanno et al., 1992).

Second, affective evaluations are closely linked with moral judgments (Haidt, 2001; Moll, Oliveira-Souza, Bramati, & Grafman, 2002). This leads to the prediction that merely understanding a behavior (thereby muting affect) can make that behavior seem more consistent with moral norms and less blameworthy — as documented in several studies. For example, when mental disorder symptoms are ordered in a coherent causal chain, people rate individuals with those symptoms as less abnormal (Ahn, Novick, & Kim, 2003; Meehl, 1973). Likewise, jurors are less likely to convict defendants when the defense can tell a coherent story using a given set of facts (Pennington & Hastie, 1992), and people are more likely to be seen as lying when they engage in unusual behaviors — even if the behaviors are irrelevant to deception (Bond et al., 1992). These findings all point to the same underlying phenomenon — when behaviors can be readily explained and meaning can be easily assigned, these behaviors are seen as more typical, more normative, and less blameworthy; conversely, when there is no available explanation for behaviors, they are seen as less normative and more blameworthy.

Third, we can ask what are the key antecedents to the feeling that an explanation is needed (e.g., Bruckmüller et al., 2017; Legare, 2012). In addition to lack of a causal chain (Ahn et al., 2003) or coherent order (Pennington & Hastie, 1992), we add perhaps the most critical antecedent of all — violation of expectations. Humans constantly predict the future and modify those predictions in light of actual events (Bar, 2007; Rescorla & Wagner, 1972). For this reason, people are strongly motivated to explain divergences from predicted behavior (Legare, Gelman, & Wellman, 2010; Wong & Yudell, 2015). As we noted earlier, people expect others to behave optimally, even when ignorant of critical information, which in turn leads people to find suboptimal behavior less readily explained than optimal behavior (Johnson & Rips, 2014).

Now we can put these ideas together. When an agent behaves suboptimally, people find that behavior difficult to explain because it violates their expectations — it does not conform to the optimal choice schema and resists attempts to make meaning of it. This feeling leads to more pronounced affective reactions to suboptimal choices, corresponding to more severe moral judgments. We thus predicted an *optimality bias* in evaluations of moral decisions, which would be mediated by the presence or absence of a coherent explanatory schema. Following previous work (Ahn et al., 2003; Johnson & Rips, 2014), we measure this explanatory gap by asking participants to indicate the extent to which they feel that an explanation is needed for the agent's behavior: If the agent behaved optimally, then participants should not feel that an explanation is needed; if the agent behaved suboptimally, then they should. These explanatory judgments should mediate the relationship between optimality and blame (as we test in Study 3).

Although this hypothesis has theoretical support, it has not been tested. The most closely related studies are the many demonstrations of the *outcome bias* (e.g., Baron & Hershey, 1988; see also Martin & Cushman, 2016a), wherein people blame agents for negative *consequences* despite positive intentions. Both biases are examples of *behaviorist* moral judgment that ignore the moral agent's mental states, but in quite different ways: The outcome bias assigns blame for bad *consequences* (holding the choice constant), whereas the optimality bias assigns blame for suboptimal *choices* (holding the consequences constant). These biases could operate independently. For example, a doctor might be blamed if a patient dies, even if she made the best possible choice (outcome bias with no optimality bias). On the other hand, a doctor who unknowingly made a suboptimal choice could be blamed even if the patient's outcome is fine (optimality bias with no outcome bias; see Study 4). The biases can also work together — if the patient dies, the doctor could be blamed *both* because of the bad outcome *and* because she unknowingly made a suboptimal choice. These possibilities are not just theoretical — juries must make such decisions everyday in malpractice suits.

It is important to subject this hypothesis to empirical test, not only because of its practical importance, but because there are theoretical reasons that an optimality bias might not occur in moral judgment. Moral judgments are other-focused rather than self-focused, yet the optimality bias in conventional judgments is stronger for predicting one's own choices (Johnson & Rips, under review, Experiment 2). Moral judgments usually occur in response to harm, which may strongly motivate blame judgments regardless of the agent's choice (Alicke, 1992; Martin & Cushman, 2016b). Indeed, moral choices differ from conventional choices on many key dimensions (e.g., seriousness, generality, authority-independence, and objectiveness; Kumar, 2015; Turiel, 1983), which have led some to posit domain-specific processes in moral cognition (Cosmides, 1989; Haidt & Joseph, 2004). For these reasons, Johnson and Rips (2014, 2015) explicitly avoided moral stimuli so as not to mix such critically distinct scenarios as medical malpractice and shampoo shopping.

Finding an optimality bias would inform key debates in moral psychology. First, researchers disagree over the relative importance of utilitarian (optimizing the happiness of individuals) versus deontic considerations (following moral rules) (e.g., Shenhav & Greene, 2010; Siegel, Crockett, & Dolan, 2017). Although utilitarianism is consistent with a difference between optimal and suboptimal outcomes, it would predict differences across suboptimal outcomes too (i.e., a 50% chance of a good outcome has higher expected utility than a 30% chance, even if a 70% chance would be even better). Because numerical stimuli tend to prompt utilitarian judgments (Shenhav & Greene, 2010), an optimality bias would demonstrate a violation of utilitarian logic on its home turf. Second, demonstrating the Efficiency Principle in moral judgment would complement recent demonstrations that moral psychology depends in part upon more domain-general mechanisms (Greene, 2015), including heuristic processes (Sunstein, 2005). Third, use of the Efficiency Principle would have important implications for debates over the role of theory-of-mind in moral judgment (e.g., Cushman, 2008), since efficiency-based thinking appears to have different psychological properties compared to fuller, representational theory-of-mind (Gergely & Csibra, 2003). We return to these and other theoretical implications in the General Discussion.

The current studies. Across seven studies, we test the existence and causes of the optimality bias in moral judgment. These studies focus on agents making morally laden decisions in which three potential options differ in quality, but agents falsely believe that the options are equivalent. If participants fall prey to the optimality bias, they would assign blame based on the quality of the agents' choices, even though the agents are ignorant. Further, if this thinking is truly based on efficiency rather than mere probability, we should expect participants to give more lenient moral judgments only if an agent makes an optimal choice; we should not expect moral judgments to differ between suboptimal choices that vary equally in quality.

Study 1 provides an initial test of this prediction for judgments of wrongness and punishment. Study 2 seeks to extend the optimality bias to cases in which the agent's ignorance is both salient and clearly justified. Study 3 provides direct evidence for the mechanism, by measuring both need for explanation and blame, and then testing our mediation model. Next, we assess the plausibility of alternative accounts by testing positive outcomes (Study 4) and by measuring two potential moderators — the agent's perceived negligence and the extent to which participants erroneously attribute knowledge to the agent (Study 5). Finally, we examine potential boundary conditions, testing whether the optimality bias persists when the agent knows the probabilities (Study 6) and when participants forecast their blame judgments (Study 7).

Study 1: Wrongness and Punishment

Study 1 examined the optimality bias in judgments of wrongness and punishment. On standard accounts of these judgments (Cushman, 2008), wrongness primarily tracks the negativity of an agent's intentions, and punishment primarily tracks the negativity of the outcome caused by the agent. Could both of these judgments also be affected by the *optimality* of an agent's choice, irrespective of the agent's knowledge?

Methods. Participants in all studies were American, and were recruited and compensated using Amazon Mechanical Turk. Relative to traditional samples of undergraduates, Mechanical Turk participants tend to be somewhat older and more highly educated, though with high variance (e.g., Paolacci & Chandler, 2014). Thus, these samples would generalize more readily to the American population compared to undergraduate samples, although we cannot make statements about cross-cultural stability. Participants provided informed consent in accordance with the procedures of the Yale University Human Subjects Committee. For each study, participants had not participated in any of the other studies. Sample sizes were planned *a priori* to achieve at least 90% power based on effect size estimates from related studies, before exclusions (for more details, see Appendix S4 in the online Supplementary Materials). In these studies, we report all measures, manipulations, and exclusions.

For Study 1, we recruited 336 participants (M_{age} = 31, 41% female); 80 were excluded due to incorrect answers to check questions (see Appendix S5 for exclusion criteria). Previous research found that excluding participants who failed comprehension checks can reduce noise in responses due to inattention in online studies (Thomas & Clifford, 2017), and our exclusion percentages were within the range of published exclusions for data collected on Mechanical Turk (e.g., Thomas et al., 2016; De Freitas et al., 2017). The conclusions of significance tests in this article did not generally depend on the exclusion criteria, with analyses repeated on the full sample leading to similar results in nearly all cases (see Appendix S5).

Participants were randomly assigned to one of eight vignettes concerning different moral agents (doctor, farmer, contractor, programmer, pilot, paramedic, CEO, or broker). Vignettes were normed on Kumar's (2015) distinctly moral attributes (seriousness, generality, authority-independence, and objectiveness) to verify their moral significance, relative to stimuli used in related work (see supplementary Study S1).

Participants judged agents who made moral decisions under uncertainty, which always led to a negative outcome. These agents always had three possible options, having a 70%, 50%, and 30% probability of leading to a favorable outcome (we refer to these as *Best*, *Middle*, and *Worst*, respectively). The agent always falsely believed, however, that the three options were of the same quality. For example, in the *Best* condition the agent behaved optimally, choosing the option with the highest probability of a positive outcome:

> A doctor working in a hospital has a patient who is having hearing problems. This patient has three, and only three, treatment options. The doctor believes that all treatment options have a 70% chance of giving the patient a full, successful recovery. But in fact the doctor's belief is wrong. Actually:
>
> 1) If she gives the patient treatment LPN, there is a 70% chance the patient will have a full recovery.
> 2) If she gives the patient treatment PTY, there is a 50% chance the patient will have a full recovery.
> 3) If she gives the patient treatment NRW, there is a 30% chance the patient will have a full recovery.
>
> The doctor chooses treatment LPN, and the patient does not recover at all. The patient now has permanent hearing loss.

The *Middle* and *Worst* conditions differed only in which choice the agent made (i.e., PTY or NRW rather than LPN). Note that the probabilistic difference between Best and Middle is the same as between Middle and Worst, but only Best maximizes the probability of the outcome. That is, Best is the optimal decision, even though the agent has no way of knowing. (See Appendix S1 for the text of other vignettes, Appendix S2 for how wording varied across studies, and Appendix S3 for a summary of differences across studies.)

On the same page, participants in Study 1A answered a question about wrongness (e.g., "How wrong was the doctor's behavior?") and participants in Study 1B answered a question about punishment ("How much should the doctor be punished?"), on a scale anchored at 1 ("not at all"), 4 ("somewhat"), and 7 ("very much"). The dependent measures were reverse-coded for consistency with other studies.

On the next page, participants answered two check questions, to ensure comprehension of the vignette. To be included, a participant had to correctly indicate the probability of success given the agent's choice and to acknowledge that the agent had a false belief about the probabilities (see Appendix S5 for wordings).

Results and discussion. Even though the agent thought that the three options were of the same quality, participants sharply differed in their moral judgments depending on the agent's choice (Figure 1). In Study 1A, agents who chose the (optimal) Best option were judged as behaving less wrongly than those who chose the (suboptimal) Middle option [$t(87) = 5.22$, $p < .001$, $d = 1.11$, 95% CI[1.24,2.76], $BF_{10} > 1000$, $d_S = 0.61$].[1] Thus, participants based their judgments of wrongness on the quality of the agent's choice, even if the agent had no way of knowing that choice quality.

However, judgments of wrongness did not simply track the probability of the outcome. Wrongness judgments were *no less harsh* when agents chose the Middle rather than the Worst option [$t(86) = -0.28$, $p = .78$, $d = -0.06$, 95% CI[–0.93,0.70], $BF_{01} = 5.9$, $d_S = 0.59$], even though the probabilistic difference between

[1] We supplement all *t*-tests reported in this paper with three pieces of information: (1) a 95% CI on the condition difference being tested. (2) A Bayes Factor (*BF*) (Rouder et al., 2009). For example, "BF_{10}=4.0" means that the data would be 4 times likelier under the alternative hypothesis than under the null hypothesis, giving reason to reject the null hypothesis. However, *BF*s can also quantify evidence in favor of a null hypothesis; "BF_{01}=6.0" means that the data would be 6 times likelier under the null than under the alternative, giving reason to accept the null hypothesis. The Bayesian analysis assumes JZS priors (Rouder et al., 2009), with a scale factor of 1. (3) A sensitivity power analysis, reporting the minimum effect size (d_S) that would be detectable with 80% power, based on the actual sample size after exclusions. These sensitivity analyses generally found that the studies were well-powered to detect a medium-sized effect. In addition, we report a meta-analysis (following Study 5) pooling data across studies with nearly 2000 participants.

Best and Middle (70% vs. 50%) was the same as that between Middle and Worst (50% vs. 30%). Participants did not simply view *worse* choices as morally wrong, but failing to choose the *very best* option.

In Study 1B, participants again wished to punish the agent to differing degrees depending on their choice. Agents who chose the Best option were judged less deserving of punishment than those who chose the Middle option [$t(74) = 3.71$, $p < .001$, $d = 0.85$, 95% CI[0.60,1.98], $BF_{10} = 64.8$, $d_S = 0.64$]. Mirroring judgments of wrongness, however, punishment judgments did not differ between agents who chose Middle and Worst [$t(84) = 1.61$, $p = .11$, $d = 0.35$, 95% CI[–0.14,1.30], $BF_{01} = 1.8$, $d_S = 0.66$]. Thus, punishment judgments too are affected not only by the negativity of the outcome caused by the agent, but also by the optimality of the agent's choice. Once again, the pattern of judgments showed a strong effect of optimality (the Best–Middle difference) but little residual effect of probability (the Middle–Worst difference).

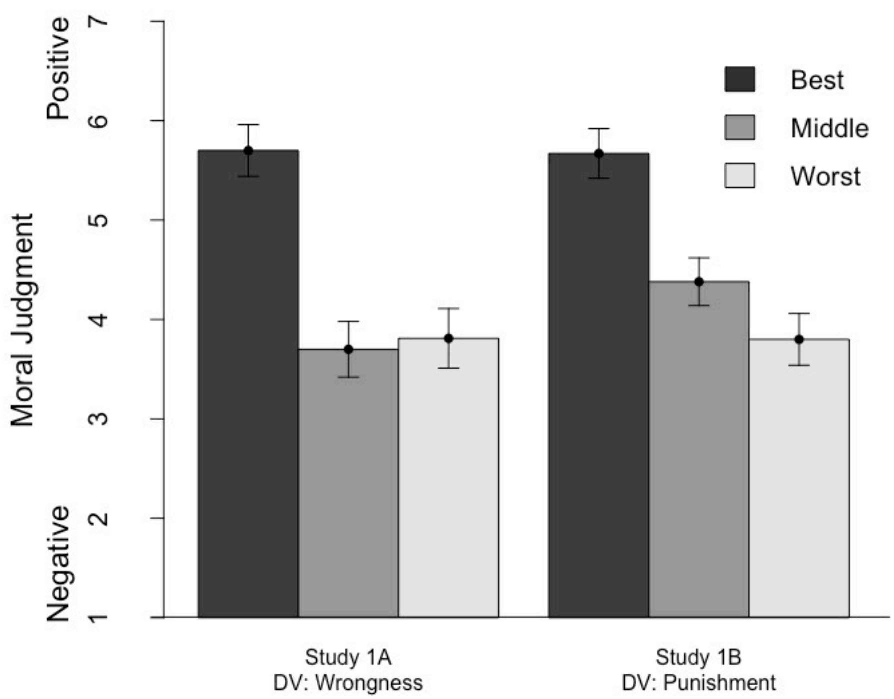

Figure 1. Results of Study 1.

Note. Bars represent 1 SE. Scales reverse-coded.

These results show that for two different moral judgments, people make more lenient judgments for agents making optimal decisions, even if they have no way of knowing which decision is optimal. These findings provide evidence for a *moral optimality bias* — a tendency to base moral judgments on the optimality of a decision over-and-above the agent's mental states. Further, people do not make more lenient evaluations for better rather than worse suboptimal decisions, but regard all suboptimal decisions as equally wrong and punishable, in accordance with the Efficiency Principle (Dennett, 1987; Gergely & Csibra, 2003).

However, various concerns about this result are possible. First, participants could have assigned culpability not for making the wrong choice, but for being ignorant. A doctor, for example, might be deemed negligent if she does not know the risks associated with various procedures. This cannot fully account for our findings, because the doctor was ignorant *even when she chose the Best option*. Nonetheless, we would expect the optimality bias to generalize to cases in which the agent could not reasonably be expected to know the risks associated with the different options. Studies 2 and 5 address this issue.

Second, participants could have been making judgments about competence rather than morality. The word "wrong" aggravates this issue, as it applies equally to immoral and incompetent choices. This concern is mitigated by the similarity between punishment and wrongness judgments and by participants' beliefs that the scenarios reflect distinctly moral behaviors (see supplementary Study S1). Nonetheless, it would be more convincing yet if the optimality bias turned up in other indices of moral judgment, such as judgments of blame or character, which more clearly track moral evaluations. Although we would expect *a priori* that the optimality bias would track all of these measures (see Cushman, 2008 for evidence that they are impacted by similar factors), we generalize these results further to judgments of blame in subsequent studies.

Study 2: Salient Ignorance about Unknowable Risks

Study 2 sought to buttress the findings of Study 1 in two primary ways. First, decision situations in real life are often associated not only with probabilistic risk (i.e., a specific probability is associated with each option) but also with *ignorance about those risks themselves* (see Knight, 1921 on a related distinction in economics). In such situations, one may be not only ignorant about the risks, but *necessarily* ignorant. A pharmaceutical company testing a new drug may justifiably have tremendous uncertainty about the probable effects of an experimental treatment; an economic policy-maker may be poorly equipped to assess the effects of various possible changes to regulatory structure or fiscal policy. We simulated such situations in our studies by specifying that the risks were *unknowable*. Thus, for reasons entirely out of their control, agents had false beliefs about the risks of each option. In addition to generalizing to such cases of radical or Knightian uncertainty, this change addresses the concern from Study 1 that participants may have been blaming agents for their ignorance itself. When the risks are in fact unknowable despite agents' due diligence, the agents could not plausibly be held as negligent.

Second, in the real world, a moral agent's ignorance is often on full display and at the center of moral debate. For instance, during the L'Aquila earthquake incident, the impossibility of predicting earthquakes was generally conceded. Nonetheless, people seem to continue to use states of the world to which the agent does not have epistemic access, even when that lack of access is salient. Study 2 tested this possibility by highlighting the agent's lack of knowledge (and the impossibility of such knowledge) *prior* to asking participants to make a blame judgment. We did this by asking participants to complete multiple choice check questions about the agent's state of knowledge before the blame measure rather than after. This manipulation should focus participants' attention on the agent's lack of access to the relevant information. Nonetheless, we predicted that participants' moral judgments would be affected by the choice's optimality.

Methods. We recruited 329 participants (M_{age} = 33, 49% female); 39 were excluded due to incorrect answers to check questions. The procedure was similar to Study 1, except for the following changes. First, the vignette text was altered to include stipulations that the agents had done due diligence and that their beliefs were rational given the available evidence. For example:

> A doctor working in a hospital has a patient who is having hearing problems. This patient has three, and only three, treatment options. Based on many articles that the doctor has carefully read in respected medical journals, she truly believes that all three options have a 70% chance of giving the patient a full, successful recovery. In fact, all of the existing evidence says that this belief is correct. But as it happens, for reasons completely outside of her control, the doctor's belief is wrong. Actually:
>
> 1) If she gives the patient treatment LPN, there is a 70% chance the patient will have a full recovery.
> 2) If she gives the patient treatment PTY, there is a 50% chance the patient will have a full recovery.
> 3) If she gives the patient treatment NRW, there is a 30% chance the patient will have a full recovery.
>
> The doctor chooses treatment LPN, and the patient does not recover at all. The patient now has permanent hearing loss.

Second, the check questions were included on the same page as the vignette. An additional check question was included about the knowability of the probabilities; participants were included only if they acknowledged that the agent's false belief was outside of her control (see Appendix S5). Third, the moral judgments were then made on the following page, with the vignette reproduced at the top of the page. The dependent measure was blame rather than wrongness or punishment, and was measured on a scale from 1 ("extreme blame") to 9 ("extreme praise"), with the vignette reproduced at the top of the page.

Results and discussion. Even though participants had just acknowledged explicitly that the agent did not know—and *could not know*—the probabilities, they nonetheless blamed agents in accordance with the Efficiency Principle. Blame judgments were more favorable for agents choosing Best rather than Middle [$t(195) = 2.11$, $p = .036$, $d = 0.30$, 95% CI[0.03,0.74], $BF_{01} = 1.1$, $d_S = 0.40$; Figure 2], but did not differ between Middle and Worst [$t(190) = 1.16$, $p = .25$, $d = 0.17$, 95% CI[–0.15,0.59], $BF_{01} = 4.6$, $d_S = 0.40$].

Coupled with Study 1, these results indicate a highly robust moral optimality bias. Participants continued to hold agents more responsible after choosing suboptimally, despite (a) the stipulation in the vignette that the probabilities were both unknown and *unknowable*; (b) including in the analysis only participants who acknowledged these stipulations; and (c) these acknowledgements being highly salient to participants when making their judgments, since they were made immediately prior to the moral judgments.

To what extent would these results generalize to the real world? One concern may be the stipulation in the vignettes that the probabilities were unknowable. Although the internal validity of the experiment requires strong constraints on the agent's ability to learn the true probabilities, we seldom face such strong stipulations in the real world. That said, unknowable *information* is ubiquitous (Knight, 1921). Even in the relatively well-defined domain of medical diagnosis, for example, doctors often face novel treatments, novel combinations of symptoms, and novel complications. Thus, although explicit stipulations may not often be encountered in real life, their underlying reality is all too common.

Two other sources of evidence suggest that the optimality bias may be quite general. First, several supplementary studies systematically vary features of the vignettes, including outcome (positive, negative, or unknown; Studies S3 and S6), intention (positive, neutral, or negative; Study S4), and knowability of the probabilities (knowable or unknowable; Study S5). In all cases, we see a substantial difference between the Best and Middle conditions and no significant difference between the Middle and Worst conditions — precisely the optimality pattern found in Studies 1 and 2 (see Figure S1). Moreover, these factors did not moderate the optimality bias. This suggests that the optimality effect is robust across many situations.

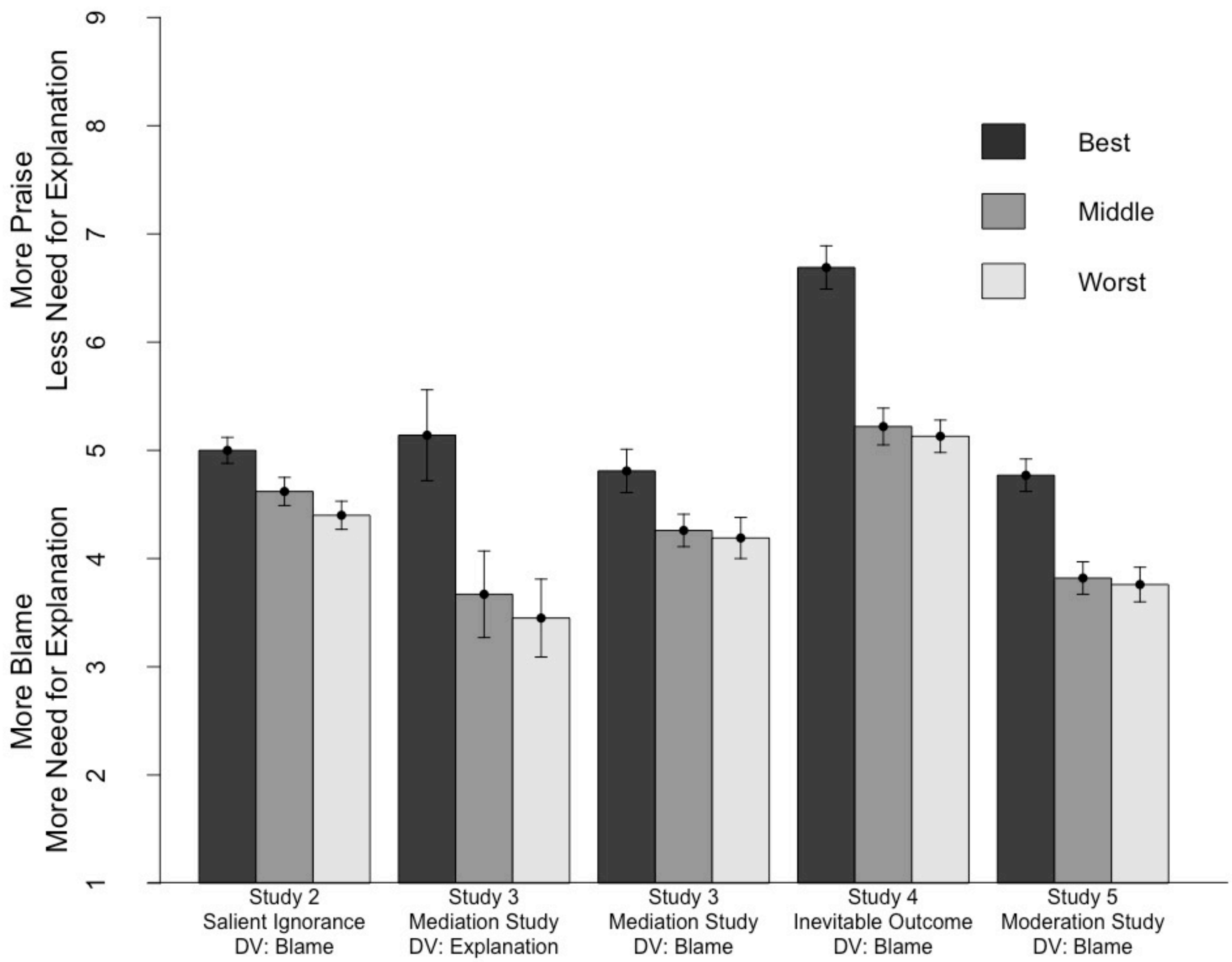

Figure 2. Results of Studies 2–5.

Note. Bars represent 1 SE. Scales reverse-coded.

Second, the findings were similar across the vignettes. The vignettes varied in the directness of the harm and the number of individuals harmed, and to some extent in the severity of the harm (see Table S1). In supplementary Study S2, we report analyses in which we try to explain differences in the magnitude of the optimality bias across vignettes (collapsing data across all between-subjects studies). This analysis revealed that the optimality bias occurred consistently for all eight vignettes, with highly significant differences between Best and Middle (see also our internal meta-analysis after Study 5). The differences between Middle and Worst were consistently smaller and only approached statistical significance for one vignette. Moreover, when we analyzed the relationship between these effect sizes and variability in the directness, severity, and number of individuals harmed, there was no statistically reliable relationship. Thus, the optimality bias appears to generalize well across very different moral contexts.

Study 3: Need for Explanation as the Mechanism

We predicted the moral optimality bias based on previous demonstrations of efficiency-based thinking about non-moral decisions. People predict that others will behave optimally and, therefore, find it more difficult to make sense of suboptimal behaviors, because such behaviors violate expectations and elude their (over-extended) explanatory schemas. That is, even when an agent is ignorant about the relevant probabilities needed to make a decision, people find optimal choices to be less in need of an explanation compared to suboptimal choices (Johnson & Rips, 2014). Based on previous research on affect and moral judgment (Ahn et al., 2003; Wilson & Gilbert, 2008), we would expect more extreme affective reactions to suboptimal (therefore unexplained) behaviors, accompanied by stronger moral condemnation.

Study 3 tested this mechanism directly by measuring the extent to which participants thought that an explanation was needed for an agent's behavior (following Johnson & Rips, 2014). We predicted that suboptimal moral choices would seem more in need for explanation than optimal choices (but that different suboptimal choices would seem equally in need for explanation), and that this difference would mediate the relationship between the agent's choice and blame judgments.

Methods. We recruited 160 participants (M_{age} = 35, 55% female); 44 were excluded due to incorrect answers to check questions. The procedure was identical to Study 2, with two exceptions. First, participants answered, "To what extent do you feel that an explanation is necessary for the doctor's choice?" on a 1-to-9 scale prior to completing the blame item (on separate pages), analogous to the question used to measure need for explanation (Johnson & Rips, 2014) or difficulty of explanation (Ahn et al., 2003) in other research. This measure was reverse-coded for analysis and presentation. Second, we presented the comprehension questions after these need for explanation and blame questions, rather than before, to address the risk of potential demand characteristics associated with presenting the comprehension questions first.

Results and discussion. Consistent with previous research (Johnson & Rips, 2014), agents' decisions were considered more explainable when they chose Best rather than Middle [$t(83) = 2.54$, $p = .013$, $d = 0.55$, 95% CI[0.32,2.62], $BF_{10} = 3.0$, $d_S = 0.62$; Figure 2] but equally explainable when choosing Middle and Worst [$t(72) = 0.39$, $p = .69$, $d = 0.09$, 95% CI[–0.90,1.35], $BF_{01} = 5.2$, $d_S = 0.61$]. Replicating Studies 1 and 2, agents were also blamed less when they chose Best rather than Middle [$t(83) = 2.25$, $p = .027$, $d = 0.49$, 95% CI[0.06,1.04], $BF_{10} = 1.7$, $d_S = 0.62$] but equally when choosing Middle and Worst [$t(72) = 0.26$, $p = .79$, $d = 0.06$, 95% CI[–0.41,0.54], $BF_{01} = 5.4$, $d_S = 0.61$]. The effect size in this study ($d = 0.49$) was larger than in Study 2 ($d = 0.30$) but smaller than in Study 1 (*d*s = 1.11 and 0.85).

In preparation for the bootstrap analysis (Preacher & Kelley, 2011), condition was dummy-coded so that Best was scored as 1 and Middle and Worst were scored as 0 (we refer to this variable as *optimality*). Unstandardized regression coefficients are in the main text and standardized coefficients in Figure 3.

There were significant direct effects of optimality on both blame, $b = 0.58$, 95% CI[0.16,1.00], and need for explanation, $b = 1.56$, 95% CI[0.60,2.52]. When both optimality and need for explanation were entered into the model, need for explanation predicted blame, $b = 0.18$, 95% CI[0.11,0.26], while optimality no longer exerted a significant direct effect, $b = 0.30$, 95% CI[–0.11,0.70]. Critically, the bootstrap results indicated a significant indirect effect of optimality on blame via need for explanation, $b = 0.28$, 95% CI[0.10,0.54]. Thus, need for explanation mediates the relationship between optimality and blame (Figure 3).

This analysis is consistent with our theoretical account: Suboptimal actions elude explanation, and unexplained behaviors are seen as less normative and more blameworthy (Ahn et al., 2003; Pennington & Hastie, 1992). Some questions remain, however. First, it would be useful to further test this mechanism through experimental manipulations, such as presenting explanations for suboptimal behavior. Second, the relationship between explanation and blame is likely bidirectional — perhaps people motivated to assign blame latch onto optimality and inexplicability as justifications for blame (e.g., Alicke, 2000; Tetlock, 2002). This cannot fully explain the effect of optimality on need for explanation, since this effect is found even in non-moral contexts (Johnson & Rips, 2014), but does complicate the theoretical picture beyond the simple mediation model presented here. Finally, the mediator explained only about half of the relationship between optimality condition and blame. This leaves open the possibility that other mechanisms also partly account for the optimality bias, and it would be useful to test alternative mediators. We leave these questions to future research, instead focusing here on experimentally testing alternative accounts.

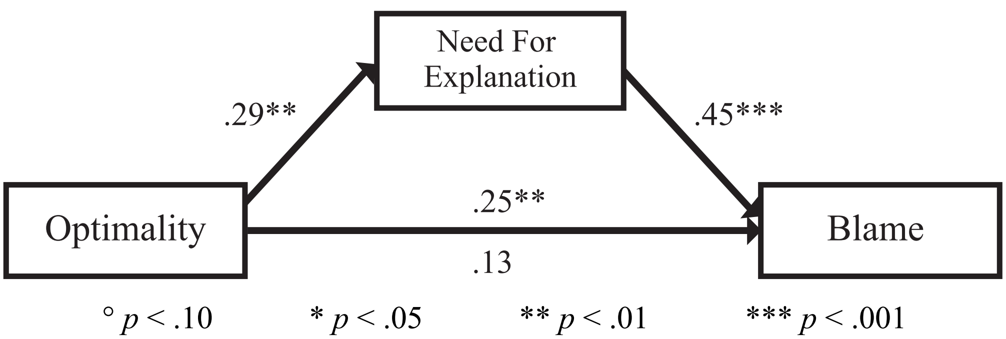

Figure 3. Mediation diagram for Study 3.

Note. Coefficients are standardized (unstandardized coefficients given in main text).

Study 4: Inevitable Positive Outcomes

Study 4 addressed alternative accounts of our findings, by distinguishing between two kinds of probability. As in previous studies, the typical probabilities associated with each option varied, and the agent either chose the Best, Middle, or Worst option. However, participants also learned about a special circumstance that rendered a positive outcome inevitable given *any* of the choices (e.g., a gene that would render any of the treatments effective). Even though the options varied in their *general* efficacy (from 70% to 50% to 30%, as in previous studies), the probability of a positive outcome in this *particular* case was always 100%, regardless of the agent's choice. Our account predicts that people should assign more negative moral judgments after a suboptimal choice, since such choices still elude explanation, even if the choice turned out not to matter in hindsight. This prediction distinguishes our account from three competitors.

First, could the results reflect differences in simulated (counterfactual) outcomes for the unrealized choices? Participants knew what happened given the agent's actual choice (always a negative outcome in previous studies) but not what would have happened given the other possible choices. In the Best condition, participants could imagine a negative outcome had the agent instead chosen Worst, recruiting downward counterfactuals (Roese, 1997) and mitigating blame; conversely, in the Worst condition, participants could imagine a positive outcome had the agent instead chosen Best, recruiting upward counterfactuals and exacerbating blame (Branscombe et al., 2003). If this accounts for the optimality bias, then the bias should be eliminated in Study 4, where the counterfactuals are specified and equated across conditions.

Second, participants could have been assuming that the probabilities were in some sense available to the agent, despite the agent's stated ignorance — that is, participants may have succumbed to *hindsight bias* (Fischhoff, 1975). One version of this concern is the notion that participants could not distinguish their own knowledge from that of the agent, an error known as the *curse of knowledge* (e.g., Birch & Bloom, 2007; Camerer, Loewenstein, & Weber, 1989) or *epistemic egocentrism* (e.g., Royzman, Cassidy, & Baron, 2003). If participants assign blame according to optimality because they tacitly imbue knowledge to the agent or otherwise believe agents should optimize their decisions in hindsight, blame should be equal across conditions in Study 4 since, in hindsight, the participant — but not the agent — knows that the true probability was 100% regardless of the agent's choice.

Third, participants could have been succumbing to outcome-based reasoning. One possibility is *culpable causation* (Alicke, 1992) — that participants carefully scrutinized agents' behavior in light of negative outcomes in order to generate reasons for blame, and then latched onto optimality as a scapegoat for motivational reasons. A second possibility is *moral luck* (Martin & Cushman, 2016b; Nagel, 1979; Williams, 1981), when an act accompanied by a positive or negative intention (e.g., helping an elderly man to cross the road, or attempting to poison one's husband) by chance has the opposite outcome (e.g., the man trips and breaks his hip, or the poison has become impotent with age). In such cases, people often will assign blame in accordance with the outcome rather than the agent's intention (Baron & Hershey, 1988; Martin & Cushman, 2016a, 2016b; see also Pizarro, Uhlmann, & Bloom, 2003). These mechanisms are theoretically compatible with our own explanation for the optimality bias. Since our studies kept the outcome constant across conditions, these outcome-based factors could at best be a necessary condition for the optimality bias. This would imply a boundary condition: The optimality bias should be eliminated when the agent has a positive intention and a positive outcome occurs, as in Study 4, since there is no negative outcome to trigger culpable causation and no mismatch between intention and outcome to generate moral luck. If these outcome-based processes operate by-and-large separately from optimality, however, the bias should persist.

Methods. We recruited 174 participants (M_{age} = 37, 55% female); 21 were excluded due to incorrect answers to check questions. The procedure for Study 4 was identical to Study 2, except for two changes. First, the outcome was positive. Second, on the page following the presentation of the initial vignette and check questions, the phrase "Here the is the story again, with new information presented in italics" was written in bold at the top of the page and the hindsight information was written at the bottom of the page in italics, indicating that the outcome would have actually been positive regardless of the choice (e.g., "We now also know that the patient had a gene that would have allowed any treatment to cure the disease"). That is, participants received probability information in two steps — first, as in previous studies, participants learned about the typical probabilities associated with the treatment options (unbeknownst to the agent); and second, unlike previous studies, participants learned about an idiosyncratic moderating factor (e.g., the gene) that rendered a positive outcome inevitable. Judgments of moral blame were made on the bottom of the second page, after the presentation of the hindsight information, using the same scale as Studies 2 and 3.

Results and discussion. In Study 4, a positive outcome inevitably occurred, but the options differed in efficiency. Thus, the Efficiency Principle predicts that agents should receive more positive moral evaluations when they choose optimally. Indeed they did. Like Studies 1–3, participants gave more positive moral evaluations when the agent chose Best rather than Middle [$t(99) = 3.38$, $p = .001$, $d = 0.67$, 95% CI[0.42,1.60], $BF_{10} = 25.6$, $d_S = 0.56$; Figure 2], but similar moral evaluations whether the agent chose Middle or Worst [$t(100) = 0.42$, $p = .67$, $d = 0.08$, 95% CI[–0.45,0.69], $BF_{01} = 6.0$, $d_S = 0.57$]. Once again, the optimality effect was large, while the residual effect of probability was negligible.

This result is compatible with our account of the optimality bias, but inconsistent with many other possibilities, including counterfactual comparison, hindsight bias, epistemic egocentrism, culpable causation, and moral luck. Indeed, supplementary Study S3 directly compares the magnitude of the optimality bias for positive and negative outcomes, and finds similar effect sizes, further speaking against the impact of culpable causation and moral luck as necessary conditions, which would predict that the effect should be smaller or eliminated for positive outcomes.

Study 5: Negligence and Egocentric Knowledge Attributions

Study 5 examines two possible moderators of the optimality bias, to further rule out two concerns raised earlier. First, people sometimes hold others accountable not in spite of their ignorance, but *because* of it, in situations of negligence. That is, agents sometimes have duties to take due diligence to ensure they know all relevant information that is available. Although Studies 2–4 stipulated that knowledge of the outcome probabilities was impossible, participants may have more tacitly assumed the possibility of such knowledge. By measuring these attributions of negligence on a continuous scale, Study 5 can test whether these more tacit attributions — held in the face of explicit stipulations — contribute to the optimality bias.

Second, participants may have tacitly imbued the agent with knowledge of the probabilities, confusing their own perspective with that of the agent. In that case, participants would have given lower blame judgments in the Middle and Worst conditions than in the Best condition because they could not fully discount their knowledge of the probabilities, even though they knew that the agents lacked that knowledge. Although this account would seem to predict no optimality bias in Study 4 (where the participant knew the true probabilities were 100%), Study 5 further tests the possible moderating influence of egocentrism on the optimality bias by measuring these egocentric attributions on a continuous scale.

Methods. We recruited 363 participants (M_{age} = 35, 57% female); 87 were excluded due to incorrect answers to check questions. The procedure was similar to Study 3, except the explanation question was not included and two moderators were measured (order counterbalanced) after the blame question, with the vignette text reproduced at the top of the screen. Measures of *negligence* ("While answering the question about blame, did you think that if the doctor had thought more carefully or done more research, then she would have been able to know which options were better and which were worse?") and *egocentrism* ("While answering the question about blame, did you think that the doctor had some sense of which options were better and which were worse?") were on scales from 1 ("not at all") to 9 ("definitely").

Results and discussion. Participants again blamed agents based on the optimality of their choices. As in Studies 1–4, blame scores were significantly less severe in the Best condition than in the Middle condition [$t(179) = 4.54$, $p < .001$, $d = 0.68$, 95% CI[0.53,1.36], $BF_{10} > 1000$, $d_S = 0.43$; Figure 5], but similar in the Middle and Worst conditions [$t(188) = 0.29$, $p = .77$, $d = 0.04$, 95% CI[–0.36,0.49], $BF_{01} = 8.5$, $d_S = 0.41$].

Our main concern was with the potentially moderating effects of negligence and egocentric knowledge attributions. To test these effects, we conducted a stepwise linear regression predicting blame ratings. Model 1 included a dummy-coded optimality variable (Best condition = '1', other conditions = '0'), as well as both moderators (centered at their means and scaled by their standard deviations). Model 1 confirmed that optimality was a large, significant predictor of blame [$b = 0.59$, $SE = 0.16$], indicating that the optimality bias operates over-and-above the (main) effects of negligence and egocentric knowledge attributions. Egocentrism also predicted blame judgments over-and-above the other predictors [$b = –0.75$, $SE = 0.08$], although negligence did not have a significant effect [$b = –0.10$, $SE = 0.08$].

Model 1 looked at only the main effects of each predictor, but we were especially interested in moderating effects. That is, was the effect of optimality *larger* for participants higher on negligence or egocentric knowledge attributions? If so, this would fuel alternative interpretations of our findings. Model 2 tested this possibility by including all two-way interaction terms as well as the three-way interaction (Table 1). None of these interactions approached significance (*p*s > .25), and the overall fit of Model 2 was not a significant improvement over Model 1 [$R^2 = .35$ vs. .34; $F(4,268) = 0.68$, $p = .60$]. Thus, attributions of negligence or egocentric knowledge do not appear to play an important role in the optimality bias.

	Model 1	**Model 2**
DV: Blame Judgments		
(Intercept)	3.91 (0.09)	3.95 (0.10)
Optimality (dummy)	0.59 (0.16) ***	0.59 (0.17) ***
Negligence	–0.10 (0.08)	–0.02 (0.10)
Egocentrism	–0.75 (0.08) ***	–0.82 (0.09) ***
Optimality x Negligence		–0.14 (0.18)
Optimality x Egocentrism		0.20 (0.18)
Negligence x Egocentrism		–0.10 (0.09)
Optimality x Negligence x Egocentrism		0.11 (0.17)
R^2	.34	.35

° $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$

$N = 276$ observations

Table 1. Tests of moderation in Experiment 5.

Note. Models of blame judgments, using as predictors a dummy-coded condition variable, continuous measures of potential negligence and egocentrism moderators (mean-centered, SD-scaled), and their interactions. Entries are regression coefficients (SEs in parentheses).

Finally, we can use the model to estimate the effect of optimality for a participant one standard deviation above or below the mean on negligence and egocentrism (by re-centering the moderators at 1 SD above or below the mean on each variable and testing the coefficient on the optimality variable). This allows us to ask whether a hypothetical participant who resisted attributions of negligence and egocentric knowledge would still show an optimality bias. The model indicates that they would indeed. A participant one SD below the mean on both negligence and egocentrism is estimated to have a significant optimality bias (i.e., significantly positive coefficient on the dummy-coded condition variable) [$b = 0.65$, $SE = 0.27$, $p = .019$], as is a participant one SD above the mean on both measures [$b = 0.77$, $SE = 0.36$, $p = .035$].

These results show that attributions of negligence and egocentric knowledge are not important moderators of the optimality bias — the effect is roughly equal in magnitude regardless of how a participant scores on these measures. This further rules out the concern that participants are assigning blame for *ignorance* itself rather than suboptimal choices. In addition, it addresses the alternative explanation that the optimality bias is driven by a curse of knowledge effect, such that participants are unable to separate their own perspective from that of the agent. Although such effects no doubt occur — indeed, blame was higher overall for participants who made more egocentric knowledge attributions — this did not interact with the optimality bias. Thus, the mechanism identified in Study 3 — the intuition of norm-violation triggered by the difficulty of explaining suboptimal actions — appears to be the primary driver of the optimality bias.

Meta-Analysis

These studies have consistently found a large and significant effect of optimality, as measured by the difference in moral judgments between the Best and Middle conditions, and a small and nonsignificant effect of probability beyond optimality, as measured by the difference in moral judgments between the Middle and Worst conditions. However, although these studies are well-powered for detecting the large optimality (Best–Middle) effect (see Appendix S4), they would be individually under-powered for detecting a smaller probability (Middle–Worst) effect. In addition, we have not attempted to directly compare the sizes of the Best–Middle and Middle–Worst differences, but only their significance levels. As the mantra goes, the difference between significant and non-significant may not itself be significant.

To address these issues, we conducted an internal meta-analysis on our between-subjects studies. This includes Studies 1–5 in the main text, as well as supplementary Studies S3A, S3B, S4A, S4B, S5A, S5B, and

S6. (Study S3C was not included because we had predicted a significant Middle–Worst difference in that study, though in fact that effect reached only marginal significance.) These studies include 1895 participants.

We conducted the meta-analysis using the METAFOR package in R (Viechtbauer, 2010). The standardized mean difference was calculated for each study, both for the Best–Middle and the Middle–Worst difference. Then, these differences were meta-analyzed using a fixed effects model. (Significance levels are robust to various random effects specifications.) Unsurprisingly, this analysis uncovered a large Best–Middle difference, consistent with the previous studies, $b = 0.72$, $SE = 0.06$, $z = 12.33$, $p < .001$, 95% CI[0.60,0.83]. This is a canonically medium to large effect, since this result is expressed in standardized units. There was also a significant Middle–Worst difference, $b = 0.14$, $SE = 0.06$, $z = 2.54$, $p = .011$, 95% CI[0.03,0.25]. Even at the high end of the confidence interval, this effect is small, but nonetheless detectable in this very large sample. Finally, since the confidence intervals on these differences do not overlap, we conclude that the Best–Middle difference is significantly larger than the Middle–Worst difference.

These meta-analytic results support the argument we have been making. First, they confirm that there is a large and statistically robust effect of optimality on moral judgment: People assign substantially less blame to agents who have made an optimal rather than a suboptimal choice. Second, the results confirm that this effect occurs over-and-above the effect of probability, since the (equally probabilistically large) gap between the agent's Middle and Worst choices corresponds to a far smaller difference in blame.

In addition, these results do suggest that there *is* some modest effect of probability over-and-above optimality, since the agents were indeed blamed more in the Worst than in the Middle condition. An interesting possibility for future work would be to test whether this effect is due to the qualitatively distinct nature of a *worst* choice. Perhaps even this small effect would be eliminated if participants were assigning blame to the second-best versus third-best choice, when there is also a fourth-best choice that is even worse.

Study 6: Knowledgeable versus Ignorant Agents

From a rational perspective, it is surprising that people would blame ignorant agents for making suboptimal choices, since these agents did not intend to choose a suboptimal choice and had no principled way to make a better choice. On the other hand, it would be surprising if people did *not* blame *knowledgeable* agents for making suboptimal choices. That is, if a doctor *knowingly* chooses a treatment that fails to maximize the patient's chance of recovery, then the doctor is reasonably regarded as culpable for a bad outcome. Study 6 directly compares the effect of optimal and suboptimal choices for knowledgeable and ignorant agents. This is important for generalizing our results to the real world, since an agent's knowledge is often ambiguous. Blame attributions for clearly knowledgeable agents should produce an upper-bound for ambiguous cases, while blame attributions for clearly ignorant agents should produce a lower-bound.

This design also allows us the opportunity to test a further prediction of our process account. We have found so far that participants do rely on probability in assigning blame, over-and-above optimality, albeit to a very small degree that is undetectable in individual studies. This was revealed by our meta-analysis that found a small difference between the Middle and Worst conditions when pooling the data across studies. However, in past work examining explanations and predictions of behavior (Johnson & Rips, 2014), people robustly relied on probability when making inferences about knowledgeable agents. That is, the choices of agents who *did* know about the probabilities were seen as *much* more in need for explanation given the Middle than the Best choice (i.e., an effect of *optimality*), but also as somewhat more in need for explanation given the Worst than the Middle choice (i.e., an effect of *probability*). Given our finding that need for explanation mediates blame judgments, we would predict that blame judgments for knowledgeable agents should follow a similar pattern: Much more severe blame for the Middle choice than for the Best choice, but also somewhat more severe blame for the Worst choice than for the Middle choice.

Study 6 adopts a within-subjects design, with each participant responding to all three choice conditions for either ignorant agents (Study 6A) or for knowledgeable agents (Study 6B). In addition to comparing judgments for different types of agents, this allows us to test whether efficiency-based thinking would hold up in a design which allows participants to compare their responses for internal consistency.

Methods. We recruited 191 participants (M_{age} = 35, 40% female); 51 were excluded due to incorrect answers to check questions. The vignettes were the same as in Study 5, except that two vignettes (jet pilot and paramedic) were randomly omitted to facilitate counterbalancing. In Study 6A, the vignettes specified that the agents did not know these probabilities, as in previous studies, whereas in Study 6B, the vignettes specified that the agent *did* know the probabilities. Each participant read three different vignettes in a within-subjects design, one each where the agent chose Best, Middle, and Worst. The order of the items was randomized, and the assignment of vignette to choice condition was counterbalanced. For each vignette, participants answered both a question about blame (similar to previous studies) and wrongness (similar to Study 1A). The dependent measures in Studies 6 and 7 were coded for consistency with the other studies.

Results and discussion. The results were similar for wrongness and blame, so we collapsed across these variables for analysis (see Figure 4 for the means broken down by measure).

The results for ignorant agents in Study 6A were similar to those of previous studies. Agents who chose Best were deemed less culpable than agents who chose Middle [$t(63) = 4.56$, $p < .001$, $d = 0.70$, 95% CI[0.65,1.65], $BF_{10} = 699.3$, $d_S = 0.36$], replicating the optimality bias. Yet, agents who chose Middle were only blamed marginally less than those who chose Worst [$t(63) = 1.87$, $p = .066$, $d = 0.21$, 95% CI[–0.02,0.73], $BF_{01} = 1.9$, $d_S = 0.36$]. Similar to previous studies (as quantified by the meta-analysis), the Best–Middle difference was significantly larger than the Middle–Worst difference [$t(63) = 2.26$, $p = .027$, $d = 0.45$, 95% CI[0.09,1.50], $BF_{10} = 1.1$, $d_S = 0.36$]. Thus, even when participants experience all three conditions of the experiment, they continue to apply the Efficiency Principle and produce an optimality bias.

The results for knowledgeable agents in Study 6B were markedly different. Agents who chose Best were judged far more leniently than those who chose Middle [$t(80) = 15.73$, $p < .001$, $d = 2.46$, 95% CI[3.12,4.02], $BF_{10} > 1000$, $d_S = 0.32$]. This optimality effect was far larger than those found in previous studies, perhaps not surprisingly since those who chose Middle were knowingly choosing an inferior option. Of greater interest, participants also gave significantly harsher judgments to agents choosing Worst rather than Middle [$t(80) = 6.38$, $p <.001$, $d = 0.88$, 95% CI[0.76,1.46], $BF_{10} > 1000$, $d_S = 0.32$]. Unlike previous studies of ignorant agents where this effect was so small as to be undetectable in an individual study, it was of large magnitude for knowledgeable agents. However, this effect of probability (Middle–Worst) was significantly smaller than the optimality effect (Best–Middle) [$t(80) = 7.10$, $p < .001$, $d = 1.35$, 95% CI[1.77,3.15], $BF_{10} > 1000$, $d_S = 0.32$]. This difference in effect sizes is consistent with the differences in need for explanation found for knowledgeable agents in prior work (Johnson & Rips, 2014), confirming a prediction of our process account. Both differences (Best–Middle and Middle–Worst) were significantly larger for the knowledgeable agents of Study 6B than for the ignorant agents of Study 6A [$t(143) = 7.13$, $p < .001$, $d = 1.19$, 95% CI[1.75,3.09], $BF_{10} > 1000$, $d_S = 0.36$ and $t(143) = 2.95$, $p = .004$, $d = 0.49$, 95% CI[0.25,1.27], $BF_{10} = 7.5$, , $d_S = 0.36$, respectively]. These differences in the response pattern across Experiments 6A and 6B led to a significant interaction between experiment and condition [$F(2,286) = 56.11$, $p < .001$].

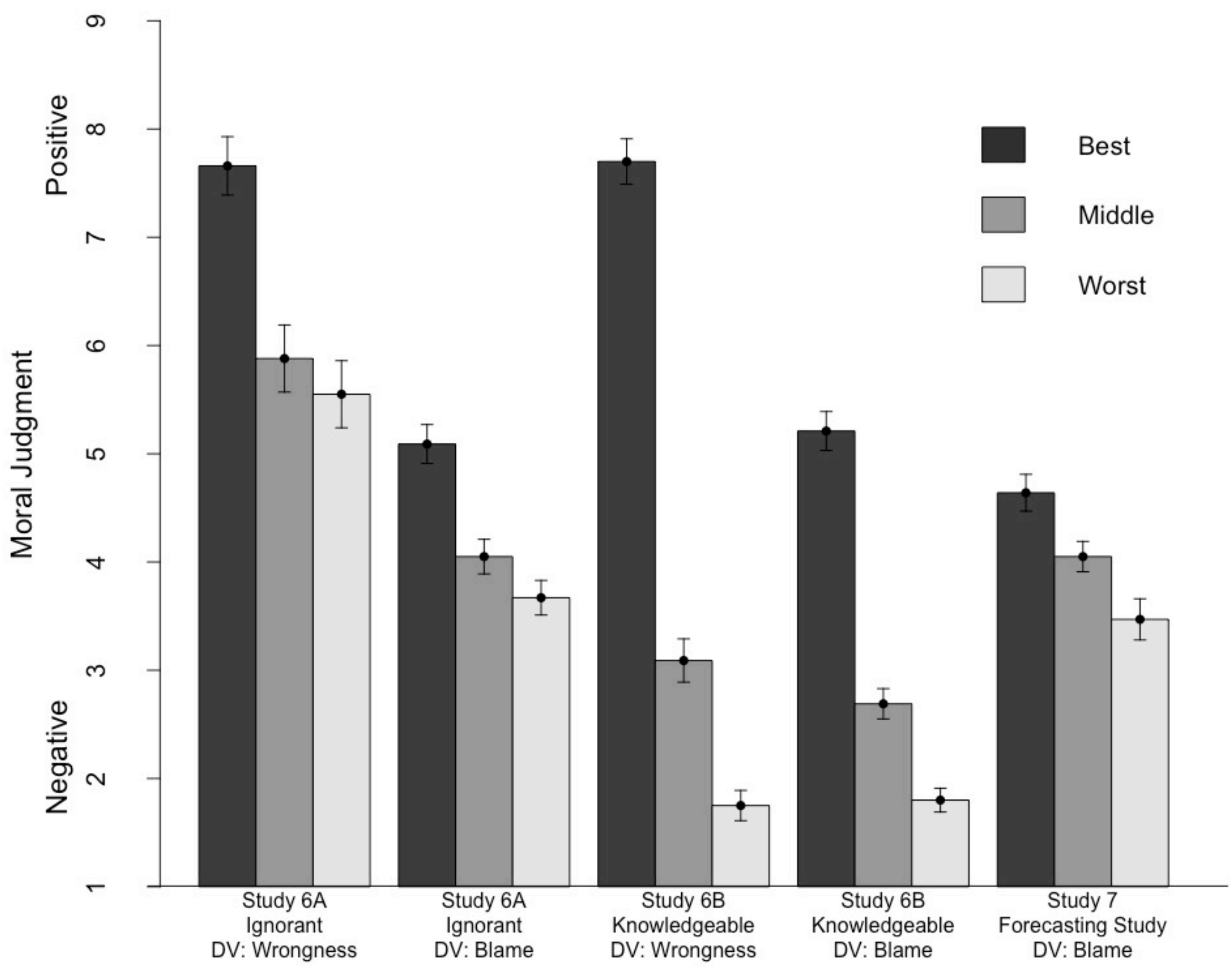

Figure 4. Results of Studies 6 and 7.

Note. Bars represent 1 SE. Scales reverse-coded.

Overall, Study 6 has three take-aways. First, participants' moral judgments are sensitive both to optimality and to probability for knowledgeable agents, and both effects are highly robust. (The effect of probability for ignorant agents, in contrast, is extremely small and difficult to detect.) This confirms a prediction made by our mediation model, since previous work has found that people use both optimality and probability when explaining (non-moral) choices (Johnson & Rips, 2014). Second, the effect of optimality for knowledgeable agents is larger than the effect of mere probability, with a far larger Best–Middle difference than Middle–Worst difference. This is again consistent with predictions, since the explainability gap between Best and Middle choices is larger than the explainability gap between the Middle and Worst choices (Johnson & Rips, 2014). Together, these results imply that people are sensitive to the Efficiency Principle for knowledgeable as well as for ignorant agents. This is important for establishing the generality of the Efficiency Principle in moral judgment, since often people *do* have some sense of the quality of decision options, and the difference in effect sizes between knowledgeable and ignorant agents gives some sense of the bounds of the effect for realistic cases where the agent's knowledge is ambiguous.

Study 7: Are People Aware of Using the Efficiency Principle?

People often mispredict how they will behave because their intuitive theories of behavior are often incorrect (e.g., Wilson & Gilbert, 2003). Do participants hold accurate intuitive theories of their moral judgments that appeal to the Efficiency Principle? Or do they instead believe that they rely on other principles for assigning blame to ignorant moral agents? Study 7 tested this question by asking participants to predict how they *would* judge agents who made each of the three choices.

Methods. We recruited 96 participants (M_{age} = 33, 55% female); 19 were excluded due to incorrect answers to check questions. The vignettes were the same as those used in Study 6A. However, rather than judging the agent given a single choice and outcome, participants were asked how they *would* respond if the agent made each of the three possible choices and a negative outcome occurred. For each vignette, participants were asked to "Imagine that the doctor chose treatment [LPN/PTY/NRW], and that the patient did not recover at all, suffering permanent hearing loss. What would the doctor deserve to receive for her behavior?" for each of the three options of the same vignette. These three judgments were all made on the same screen, in a random order, using the same blame measure as previous studies.

Results and discussion. People appear to not be aware of using the Efficiency Principle in moral judgment, but interestingly, some participants did seem to be aware that their judgments would depend on the agent's knowledge (see "Individual Differences" below). On average, participants distinguished among all *three* options [$t(76) = 3.82$, $p < .001$, $d = 0.43$, 95% CI[0.28,0.89], $BF_{10} = 66.5$, $d_S = 0.32$ for Best versus Middle; $t(76) = 5.04$, $p < .001$, $d = 0.40$, 95% CI[0.35,0.82], $BF_{10} > 1000$, $d_S = 0.32$ for Middle versus Worst;

Figure 4]. This result suggests that people cannot accurately forecast their moral judgments about ignorant moral agents, failing to appreciate the importance of efficiency-based considerations in blame.

These results inform debates about the introspective access of moral judgment. Theories differ in the extent to which they posit inaccessible intuitions or reasoned deliberation as the driving force in moral judgment (e.g., Haidt, 2001; Pizarro & Bloom, 2003). These disagreements may occur in part because different principles of moral judgment may differ in their availability to introspection (Cushman, Young, & Hauser, 2006). For instance, people can verbalize the principle that harms caused by actions are more blameworthy than harms caused by omissions (Baron & Ritov, 1994); yet they rarely verbalize the principle that harms are more blameworthy if they are direct effects of an action rather than side effects, even though many people's judgments are in fact influenced by this distinction (Foot, 1967; Royzman & Baron, 2002). The Efficiency Principle appears to be another principle that drives moral judgment outside of awareness.

As a further test of participants' (lack of) introspective access, supplementary Study S6 asked participants to justify their response patterns, allowing participants to choose a justification for their moral judgments that either explicitly acknowledged the exculpating quality of the agent's ignorance (a mentalistic justification) or which held that the agent's ignorance did not excuse their choice (an efficiency-based justification). Not only did two-thirds of participants choose the mentalistic justification, but participants had an equally large optimality bias regardless of their justification. People appear quite unaware of the extent to which the Efficiency Principle drives their moral judgments.

These results also help to further rule out a possible concern about previous studies — that participants were responding to pragmatic implicatures (to use information supplied by the experimenter) or demand characteristics (to comply with the experimenter's intentions). This concern is undercut by the finding that participants' own intuitive theory of the task leads to a different response pattern than what we found in previous studies. Together with the fact that the effects were nonlinear (an unlikely pattern for an experimenter to demand) and demonstrated in several between-subjects studies (in which participants could not compare conditions), it appears unlikely that demand characteristics and pragmatic implicatures drove the effects.

Individual Differences

Because Studies 6 and 7 manipulated the agent's choice within-subjects, they allow the opportunity to examine individual patterns of responses. Table 2 summarizes the proportion of participants whose response pattern for blame judgments was either to differentiate among all three conditions ("All Different") or to give the same response for all three conditions ("All Same"). The condition means are also included for those participants who did not fall into either of these response patterns.

	Proportion of Responses		**Means for Uncategorized Responses**		
	All Different	**All Same**	**Best**	**Middle**	**Worst**
Study 6A (Ignorant)	17.2%	15.6%	4.51	3.81	3.67
Study 6B (Knowledgeable)	51.9%	3.7%	4.83	2.08	2.08
Study 7 (Forecasting)	35.1%	37.7%	3.67	3.81	3.62

Table 2. Response patterns in Studies 6 and 7.

Note. The "Proportion of Responses" columns give the proportion of responses for which either Best > Middle > Worst ("All Different") or for which Best = Middle = Worst ("All Same") for the blame judgments. The "Means for Uncategorized Responses" columns give the mean blame judgments for the participants not categorized into either of the patterns.

These analyses allow us to answer several questions about individual response patterns. First, is the optimality bias driven by a subset of participants or do most participants fall prey to this bias? In Study 6A, only 15.6% of participants gave similar ratings across the three conditions, and a further 17.2% of participants differentiated among all three options. The remaining participants could not be classified into either pattern, but these participants' responses average out to the optimality pattern, with a large gap between Best and Middle and smaller gap between Middle and Worst. Thus, the majority of participants appear to fall prey to the optimality bias. Overall, 45.3% of participants individually showed the optimality bias, differentiating between Best and Middle, with a larger Best–Middle than Middle–Worst difference.

Indeed, these analyses underestimate the proportion of participants with this underlying judgment pattern because condition was confounded with vignette for individuals (but not at the group level).

Second, do some participants think in an efficiency-based manner about knowledgeable agents? In Study 6B, more than half of participants (51.9%) differentiated among all three conditions, while only a tiny minority (3.7%) gave the same rating in all three conditions. However, the remaining participants gave responses that are efficiency-based, with a very large gap between the Best and Middle conditions, and a smaller gap between the Middle and Worst conditions. Overall, fully 64.2% of participants individually followed the optimality pattern, by the criteria used above (again, this is an underestimate). Thus, the Efficiency Principle seems to govern the blame judgments, even for knowledgeable agents.

Finally, could a subset of participants have intuited the efficiency-based pattern in the forecasting task? In Study 7, the majority of participants either differentiated among all three options (35.1%) or gave identical ratings to all three options (37.7%). Unlike Study 6A, where the remaining uncategorized participants gave responses averaging out to the efficiency pattern, the uncategorized responses in Study 7 averaged out to noise, with no systematic differences among the three options. By the criteria used above, only 19.5% of participants showed an optimality bias, and this is *not* an underestimate because vignette was not confounded with condition for individual participants in this study. This further buttresses our claim that people lack introspective awareness of using the Efficiency Principle.

General Discussion

These studies show that moral judgments are more favorable for agents who make optimal choices than for those who make suboptimal choices — even when agents are ignorant about the quality of their choices — yet the *degree* of suboptimality matters very little (Studies 1–6). This pattern occurred because suboptimal choices prevented participants from applying an optimal choice schema to the agents' behavior, making the agents' choices be seen as more in need for explanation and as correspondingly less morally normative (Study 3). This pattern persisted under conditions where other accounts would predict it should not (Studies

4 and 5), while it was *not* observed when participants were asked to forecast their judgments (Study 7), suggesting that efficiency-based moral judgment operates outside of awareness.

In what follows, we discuss what these findings tell us about the relationship between mental-state inference and moral judgment, and for several other key issues in moral psychology.

Moral behaviorism. When we judge someone's wrongdoing, we do so largely on the basis of what she was thinking. Theories of blame account for this key fact in different ways. For example, the *path model* (Malle, Guglielmo, & Monroe, 2014) holds that social perceivers assign blame based on a decision tree, where one of the key decision nodes is the agent's intentionality (see Shaver, 1985 and Weiner, 1995 for related stage-like models). In the *dual-systems* model (Cushman, 2008), all moral judgments depend fundamentally on the agent's mental states, though some types of judgments (such as wrongness) depend much less on causality. In *person-centered* models (Uhlmann, Pizarro, & Diermeier, 2015), blame judgments take account of an agent's motivation and thus inform judgments of character.

Yet, we found here that people also rely on states of the world to assign blame, and may even do so by overriding or ignoring an agent's mental states. Even when agents were ignorant about the moral decisions they faced, people blamed them more when they behaved suboptimally. Given the agent's ignorance, this suboptimal choice was contingent on the *world* itself rather than on the agent's *representation* of the world. This can be considered a species of *moral behaviorism*, in that people bypass the agent's mental states to assess blame. Since mentalizing behavior is often effortful (Lin, Keysar, & Epley, 2010), we would predict variability in the extent of behaviorist moral thinking. In particular, factors that inhibit theory of mind, either situationally or dispositionally, should exacerbate behaviorist tendencies such as the optimality bias.

Other phenomena in moral judgment can also be thought of as examples of behaviorist moral thinking. For example, in *moral luck* phenomena, people blame others for negative outcomes that were not intended, and thus the agents' intentions are ignored (Baron & Hershey, 1988; Berg-Cross, 1975; Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009; Gino, Shu, & Bazerman, 2010; Mazzocco, Alicke, & Davis, 2004; Young et al., 2007). Nonetheless, moral luck is distinct from efficiency-based thinking in several ways. Moral luck is based on outcomes (Martin & Cushman, 2016b), whereas the optimality bias can occur even when the outcome is positive (Studies 4 and S3b) or even unspecified (Study S6). Moral luck depends on the presence and number of upward counterfactuals (Martin & Cushman, 2016a), whereas we found the optimality bias even when the counterfactuals are all identical (Study 4). And moral luck tends to show up principally in judgments of punishment and blame rather than wrongness, perhaps because it is pedagogically useful to punish someone for accidentally causing a bad outcome (Martin & Cushman, 2016b; see also Cushman, 2008; Cushman, Sheketoff, Wharton, & Carey, 2013), whereas the optimality bias is equally robust for judgments of wrongness. Further studies of the potentially additive or interactive effects of outcome-based processes (such as moral luck) and efficiency-based processes (such as the optimality bias) could be a useful direction for future research.

Implications for debates in moral psychology. The use of efficiency-based thinking in moral judgment, along with other demonstrations of behaviorist moral judgment, demonstrate that moral judgment depends on mechanisms other than just mental-state inference. The nature of the efficiency-based mechanism demonstrated here is informative for several key debates in moral psychology.

First, the Efficiency Principle is *domain-general*: It is used across many different psychological faculties, not solely in moral judgment. Researchers have debated the extent to which moral judgment is rooted primarily in domain-general mechanisms that are shared across many psychological processes, versus domain-specific mechanisms that operate only in the moral domain (Cushman, 2008; De Freitas, Tobia, Newman, & Knobe, 2016; Haidt & Joseph, 2004; Mikhail, 2007; Shenhav & Greene, 2010; Turiel, 1983; see also Greene, 2015). One strategy toward addressing this problem is to test whether *particular* domain-general mechanisms are used in moral judgment, such as causal attribution (Cushman & Young, 2011), psychological essentialism (De Freitas, Cikara, Grossmann, & Schlegel, 2017; Newman, De Freitas, & Knobe, 2015; Strohminger, Knobe, & Newman, 2017), and language (Costa et al., 2014). The current work also falls in this category. People use the Efficiency Principle in other domains of cognition, including behavior prediction (Baker, Tenenbaum, & Saxe, 2009; Johnson & Rips, 2015), visual perception (Gao &

Scholl, 2011), and language understanding (Davidson, 1967; Grice, 1989). Given that this mechanism is also used in moral judgment, the current work highlights another way that moral psychology is rooted in domain-general processes, rather than reflecting the operation of a highly domain-specific moral faculty.

Second, the Efficiency Principle is a *heuristic*: In the current studies, it is an overextension of the otherwise useful rule to assume that people will behave optimally. Some aspects of moral judgment can be modelled as the output of a rational process (e.g., Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015), but moral judgments also seem to depend, at least in part, upon heuristics or rules of thumb that can sometimes lead us astray in predictable ways (e.g., Sunstein, 2005). Consistent with the latter view, our results demonstrate that people blame others for actions they had no way of knowing were suboptimal. These judgments occurred because people attempt to apply their "optimal choice" schema for making sense of behavior, and assign higher blame to the agent when that schema falls short. Although there is no universally agreed on definition of a heuristic (Chow, 2014), the optimality bias seems to qualify on at least three conceptions. It qualifies as *effort reduction* (Shah & Oppenheimer, 2008) because it is cognitively demanding and computationally intensive to use theory-of-mind, whereas behaviorist work-arounds are less effortful. It qualifies as *attribute substitution* (Kahneman & Frederick, 2002), as people substitute a cognitively accessible and readily evaluable attribute (coherence or ease of sense-making) for a more difficult judgment (normative blameworthiness). And arguably, it qualifies as *information exploitation* (Chow, 2014; Gigerenzer & Todd, 1999), in that the optimality bias exploits a cue that is usually a good guide to blameworthiness: It is only when agents are ignorant that it becomes irrational to assign blame to suboptimal moral choices. Such a simple rule may often identify violators of moral norms, but it casts too broad a net and assigns blame to those who could not have possibly known better.

Third, the Efficiency Principle is used *outside of awareness*. The role of more deliberative versus intuitive processes in moral judgments has long been debated by philosophers, and empirical methods have recently been brought to bear on this question (Cushman et al., 2006). Contrary to more unitary views, some moral principles may be more subject to conscious awareness (such as the action/omission distinction) compared to others (such as the direct/indirect distinction). The current results inform this debate by supplying another moral principle that appears to operate *outside* of awareness. Participants in Study 7 failed to predict that their blame judgments would follow an efficiency pattern, and participants in Study S6 often provided mentalistic justifications even when their responses were efficiency-based.

Fourth, the Efficiency Principle is a *deontic moral rule*. Philosophers and psychologists have disagreed over the role of *utilitarian* criteria for moral judgment and behavior (maximizing the happiness of individuals) versus *deontic* criteria (following a set of rules), with psychologists uncovering a number of moderating factors, including emotion, moral character, and egoism (Greene et al., 2001; Kahane et al., 2015, 2018; Shenhav & Greene, 2010; Siegel, Crockett, & Dolan, 2017). Because utilitarian and deontic criteria often reach similar conclusions, this debate has been difficult to adjudicate. The current findings occupy a unique position in this debate. At first glance, the Efficiency Principle may seem *highly* utilitarian — indeed, a reflexive overapplication of utilitarian standards in cases where agents could not possibly follow them — because people blame agents less for choosing an option that maximizes the victim's utility (a 70% chance of a positive outcome) over one that does not (a 50% chance). However, efficiency-based blame judgments depart sharply from utilitarianism, because such judgments are no harsher for *worse* suboptimal options (a 30% chance). Given that numerical information seems to trigger utilitarian judgments (Shenhav & Greene, 2010), our participants' use of deontic considerations contradicts utilitarian judgment in precisely the sort of situation where it ought to be at its strongest. Given that many moral judgment studies compare two options, rather than three or more where the Efficiency Principle is free to emerge, many demonstrations of utilitarian judgment may in fact be manifestations of a deeper deontic application of the Efficiency Principle.

Fifth, and finally, the Efficiency Principle is an *explanatory principle*. Previous researchers in moral psychology have not generally emphasized the importance of explanatory or sense-making processes, aside from the important role of theory of mind. We think this is an oversight. Explanatory reasoning — the set of processes people use for evaluating hypotheses in light of evidence — is emerging as a key area of interest in high-level cognition (Keil, 2006; Lombrozo, 2016), and recent work suggests that explanatory reasoning relies on a number of domain-general heuristics. For instance, people favor simpler explanations over more complex ones when evaluating causal explanations (Lombrozo, 2007) and categorizing individuals (Johnson, Kim, & Keil, 2016), and similar heuristics are even used in some visual tasks (Johnson, Jin, & Keil, 2014). Likewise, people favor explanations that do not make unverified predictions in causal reasoning and categorization (Johnson, Rajeev-Kumar, & Keil, 2016; Khemlani, Sussman, & Oppenheimer, 2011; Sussman, Khemlani, & Oppenheimer, 2014) as well as some decision-making contexts (Johnson, Zhang, & Keil, 2016). If the relationship between explanation and moral reasoning generalizes beyond the current studies — and we have no reason to think it would not — this suggests a variety of roads for new empirical and theoretical work at the intersection of moral judgment and high-level cognition.

Practical implications. It is often difficult to predict the consequences of one's actions, and so one may sometimes make choices that turn out, in retrospect, to be suboptimal. In such situations, these results point to a risk that observers will blame actors who are trying their best under conditions of ignorance.

Legal courts are often responsible for assessing culpability when a defendant was ignorant of some important aspects of a situation, as in many cases of medical malpractice (Raghuveer, 2015). Indeed, doctors are often faced with the prospect of treating children whose parents *demand* suboptimal treatments (Nair et al., 2017). Such situations place jurors in inherently difficult situations, which are compounded by biases in moral judgment such as the optimality and outcome biases. Studying these biases in applied settings such as juror decision-making could have great practical import (cf. Pennington & Hastie, 1992).

Likewise, in our lives as consumers, we often experience moral outrage at the behaviors of companies (Antonetti & Maklan, 2016), which often devote tremendous resources to minimizing risk to their customers but nonetheless are not omniscient. Inevitably, some medications will have unintended side-effects, some cars will have defects, and some employees will act out. Such perceived failures of corporate social responsibility can have grave consequences for companies' bottom lines, including competitive disadvantage, consumer boycotts, and legal or regulatory actions (see Orlitzky et al., 2003 on links between profitability and corporate responsibility). Yet, it is not always clear that consumer outrage is fair (e.g., when consumers proposed boycotting Olive Garden restaurants in 2013 simply because the chain's parent company is located in Florida, where the controversial Zimmerman verdict was handed down; Tuttle, 2013). Our findings suggest that ignorance may not be a suitable line of defense in such cases; testing alternative communications strategies for consumer appeasement may be useful.

As for the Italian scientists who failed to predict an earthquake, it seems likely that the outrage was fueled by rationalizations, built upon judgments ultimately driven by an efficiency-based heuristic. This cognitive bias account stands in contrast to many other narratives surrounding the case. Some commentators argued that the proceedings reflected Italy's contempt for scientists (Nature, 2012), or that it was the Italian government's attempt to find a scapegoat (Hall, 2011). For their part, the families of the victims — and much of the scientific community — said that the sentence simply did not make any sense (Nature, 2012). To create justice in a world of both morally fallible actors and cognitively fallible observers, we must first look inward to understand our own biases.

Acknowledgements

Studies 1–3 were presented at the 37th Annual Conference of the Cognitive Science Society. We thank Fiery Cushman and Justin Martin for helpful comments.

References

Ahn, W., Novick, L. R., & Kim, N. S. (2003). Understanding behavior makes it more normal. *Psychonomic Bulletin & Review, 10,* 746–752.

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.

Antonetti, P., & Maklan, S. (2016). An extended model of moral outrage at corporate social irresponsibility. *Journal of Business Ethics*, *135*, 429–444.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition 113*, 329–349.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*, 280–289.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality & Social Psychology, 54*, 569–579.

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

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral judgment and decision making. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making*. Chichester, UK: Wiley-Blackwell.

Berg-Cross, L. (1975). Intentionality, degree of damage, and moral judgments. *Child Development, 46*, 970–974.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*, 382–386.

Bonanno, G. A., Wortman, C. B., Lehman, D. R., Tweed, R. G., Haring, M., Sonnega, J., et al. (2002). Resilience to loss and chronic grief: A prospective study from preloss to 18-months postloss. *Journal of Personality and Social Psychology, 83,* 1150–1164.

Branscombe, N. R., Wohl, M. J. A., Owen, S., Allison, J. A., & N'gbala, A. (2003). Counterfactual thinking, blame assignment, and well-being in rape victims. *Basic and Applied Social Psychology*, *25*, 265–273.

Bruckmüller, S., Hegarty, P., Teigen, K. H., Böhm, G., & Luminet, O. (2017). When do past events require explanation? Insights from social psychology. *Memory Studies*, *10*, 261–273.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*, 1232–1254.

Cartlidge, E. (2014). Updated: Appeals court overturns manslaughter convictions of six earthquake scientists. *Science News*.

Chow, S. J. (2014). Nany meanings of 'heuristic'. *The British Journal for the Philosophy of Science*, *66*, 977–1016.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*, 187–276.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS One*, *9*(4), e94842.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*, e57410.

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition, 72*, 237–267.

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

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

Cushman, F. A., Sheketoff, R., Wharton, S., & Carey, S. (2013) The development of intent-based moral judgment. *Cognition, 127*, 6–21.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*, 1052–1075.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment testing three principles of harm. *Psychological Science, 17*, 1082–1089.

Davidson, D. (1967). Truth and meaning. *Synthese, 17*, 304–323.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences, 21*, 634–636.

De Freitas, J., Sarkissian, H., Newman, G. E., Grossman, I., De Brigard, F., Luco, A., & Knobe, J. (2017). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science*. Advanced online publication.

De Freitas, J., Tobia, K., Newman, G., & Knobe, J. (2016). Normative judgments and individual essence. *Cognitive Science*, *41*, 382–402.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception & Performance*, *1*, 288–299.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception & Performance, 37*, 669–684.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature, 415*, 755–755.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences, 7*, 287–292.

Gigerenzer, G. and Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd and the ABC Research Group (Eds.), *Simple Heuristics that Make Us Smart* (pp. 3–34). New York, NY: Oxford University Press.

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless + harmless = blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*, 93–101.

Gray, K., & Wegner, D. M. (2008). The sting of intentional pain. *Psychological Science, 19*, 1260–1262.

Greene, J. D. (2015). The rise of moral cognition. *Cognition, 135*, 39–42.

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

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

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

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*, 55–66.

Hall, S. S. (2011). Scientists on trial: At fault? *Nature, 477*, 264–269.

Johnson, S. G. B., Jin, A., & Keil, F. C. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 701–706). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Kim, H. S., & Keil, F. C. (2016). Explanatory biases in social categorization. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 776–781). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology*, *89*, 39–70.

Johnson, S. G. B., & Rips, L. J. (2014). Predicting behavior from the world: Naïve behaviorism in lay decision theory. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 695–700). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology, 77*, 42–76.

Johnson, S. G. B., & Rips, L. J. (under review). *An optimality heuristic in predicting and understanding behavior.*

Johnson, S. G. B., Zhang, M., & Keil, F. C. (2016). Decision-making and biases in causal-explanatory reasoning. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1967–1972). Austin, TX: Cognitive Science Society.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensoinal model of utilitarian psychology. *Psychological Review*, *125*, 131–164.

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, UK: Cambridge University Press.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–535.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1128–1128). Austin, TX: Cognitive Science Society.

Klein, N., & Epley, N. (2014). The topography of generosity: Asymmetric evaluations of prosocial actions. *Journal of Experimental Psychology: General, 143*, 2366–2379.

Knight, F. H. (1921). *Risk, uncertainty, and profit*. Boston, MA: Hart, Schaffner, & Marx.

Kross, E., Ayduk, O., & Mischel, W. (2005). When asking "why" does not hurt: Distinguishing rumination from reflective processing of negative emotions. *Psychological Science*, *16*, 709–715.

Kumar, V. (2013). Moral judgment as a natural kind. *Philosophical Studies, 172*, 2887–2910.

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis- testing behavior in young children. *Child Development*, *83*, 173–185.

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning: Causal explanatory reasoning in children. *Child Development*, *81*, 929–944.

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.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–257.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*, 147–186.

Martin, J. W., & Cushman, F. (2016a). Why we forgive what can't be controlled. *Cognition*, *147*, 133–143.

Martin, J. W., & Cushman, F. A. (2016b). The adaptive logic of moral luck. In J. Sytsma & W. Buckwalter (Eds.), *The Blackwell Companion to Experimental Philosophy*. Chichester, UK: Wiley-Blackwell.

Mazzocco, P. J., Alicke, M., & Davis, T. L. (2004). On the robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology*, *26*, 131–146.

Meehl, P. E. (1973). *Psychodiagnosis: Selected papers*. Minneapolis, MN: University of Minnesota Press.

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

Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *NeuroImage*, *16*, 696–703.

Morgenstern, O., & von Neumann, J. (1947). *The theory of games and economic behavor.* Princeton, NJ: Princeton University Press.

Nagel, T. (1979). *Mortal questions*. New York, NY: Cambridge University Press.

Nair, T., Savulescu, J., Everett, J., Tonkens, R., & Wilkinson, D. (2017). Settling for second best: When should doctors agree to parental demands for suboptimal medical treatment? *Journal of Medical Ethics*, *43*, 831–840.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, *54*, 286–295.

Nature. (2012). Shock and law: The Italian system's contempt for its scientists is made plain by the guilty verdict in L'Aquila. *Nature*.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Belief in the true self explains asymmetries in moral judgment. *Cognitive Science, 39*, 96–125.

Nobes, G., Panagiotaki, G., & Pawson, C. (2009). The influence of negligence, intention, and outcome on children's moral judgments. *Journal of Experimental Child Psychology, 104*, 382–97.

Orlitzky, M., Schmidt, F. L., & Rynes, S. L. (2003). Corporate social and financial performance: A meta-analysis. *Organization Studies*, *24*, 403–441.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology, 62*, 189–206.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of moral intuitions: Comment on Haidt (2001). *Psychological Review*, *110*, 193–196.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*, 653–660.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods, 16*, 93–115.

Raghuveer, A. (2015). Jury awards $10.9M in malpractice case against Maumee OB-GYN. *NBC 24 News*.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York, NY: Appleton–Century–Crofts.

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Royzman, E., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*, 165–184.

Royzman, E., Cassidy, K. W., & Baron, J. (2003). "I know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology*, *7*, 38–65.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, NY: Springer.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*, 667–677.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2015). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, *12*, 551–560.

Sunstein, C.R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*, 531–570.

Sussman, A. B., Khemlani, S. S., & Oppenheimer, D. M. (2014). Latent scope bias in categorization. *Journal of Experimental Social Psychology*, *52*, 1–8.

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197.

Thomas, K. A., De Freitas, J., DeScioli, P., & Pinker, S. (2016). Recursive mentalizing and common knowledge in the bystander effect. *Journal of Experimental Psychology: General, 145*, 621–629.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, UK: Cambridge University Press.

Tuttle, B. (2013). Angered by Zimmerman verdict, some call for a boycott of Florida businesses. *Time.com*.

Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science*, *3*, 305–309.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*, 72–81.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1-48.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, NY: Guilford.

Williams, B. (1981). *Moral luck*. Cambridge, UK: Cambridge University Press.

Wilson, T. D., Centerbar, D. B., Kermer, D. A., & Gilbert, D. T. (2005). The pleasures of uncertainty: Prolonging positive moods in ways people do not anticipate. *Journal of Personality and Social Psychology*, *88*, 5–21.

Wilson, T. D., & Gilbert, D. T. (2008). Explaining away: A model of affective adaptation. *Perspectives on Psychological Science*, *3*, 370–386.

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. *Advances in Experimental Social Psychology*, *35*, 345–411.

Wong, W., & Yudell, Z. (2015). A normative account of the need for explanation. *Synthese*, *192*, 2863–2885.

Young, L., Cushman, F. A., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*, 8235–8240.

Supplementary Materials

Study S1: Norming for Moral Distinctiveness

Norming Study S1 sought to ensure that participants perceived the vignettes used in the current studies as distinctly moral, in opposition to the conventional decision-making dilemmas used in previous research (e.g., Johnson & Rips, 2015).

Methods. We recruited 49 participants (M_{age} = 36, 54% female); 22 were excluded due to incorrect answers to check questions.

Participants read thirteen vignettes in a random order — the eight vignettes used in the current studies and the five vignettes used in a previous study of lay decision theory (Johnson & Rips, 2015), such as the item concerning Jill's choice of shampoo described in the main text. After the description of each dilemma, participants were told that the agent deliberately chose the Worst option (e.g., the doctor deliberately chooses the option most likely to lead to permanent hearing loss, or Jill deliberately chooses the option least likely to make her hair smell like apples). Participants then rated the following (in this order) on scales from 1 to 9:

Seriousness. "How seriously wrong is [the doctor's] action?"

Universality. "Is it wrong in other places and times?"

Authority-independence. "Would it still be wrong if someone in authority said it was OK?"

Objectivity. "Imagine that someone else disagrees with you about whether the action was wrong. Must one of you be wrong?"

Harm. "Does the action lead to serious harm?"

Items were presented in a random order.

Results and discussion. On every measure, all of the morally laden vignettes were rated above the scale midpoint and all of the morally neutral vignettes were rated below the scale midpoint. These differences were significant for all measures: wrongness [$M = 8.48$, $SD = 0.73$ vs. $M = 3.38$, $SD = 2.03$; $t(26) = 12.23$, $p < .001$], universality [$M = 8.27$, $SD = 0.93$ vs. $M = 3.30$, $SD = 2.05$; $t(26) = 12.00$, $p < .001$], authority [$M = 7.86$, $SD = 1.61$ vs. $M = 3.21$, $SD = 2.11$; $t(26) = 9.75$, $p < .001$], objectivity [$M = 7.84$, $SD = 1.28$ vs. $M = 3.53$, $SD = 2.47$; $t(26) = 9.85$, $p < .001$], and harm [$M = 8.30$, $SD = 0.60$ vs. $M = 1.87$, $SD = 1.40$; $t(26) = 21.87$, $p < .001$]. Thus, the vignettes used in the current study were distinctly moral.

Study S2: Exploring Differences Among Vignettes

The purpose of Norming Study S2 was to test whether the seriousness of the harm invoked in each vignette — measured in terms of severity, directness, and number harmed — can explain differences in the size of the optimality bias across vignettes.

Methods. We recruited 38 participants ($M_{age} = 32$, 45% female); 4 were excluded due to incorrect answers to check questions.

Participants read the eight vignettes used in the current studies. After a description of each dilemma, participants were told that the agent deliberately chose the Worst option, as in Norming Study S1. Participants then rated the following (in this order) on scales from 1 to 9:

Severity. "How serious were the consequences of the [doctor's] actions?"

Directness. "Did the [doctor's] actions lead to direct, bodily harm?"

Number. "How many people were harmed because of the [doctor's] actions?"

Items were presented in a random order.

	Best–Middle	Middle–Worst	Severity	Directness	Number
Doctor	1.20***	0.30	8.24	7.94	1.62
Farmer	0.75*	0.19	8.44	8.24	3.97
Building Contractor	1.08***	0.21	8.68	8.68	2.79
Computer Programmer	1.42***	0.12	8.09	2.09	5.03
Jet Pilot	1.07***	0.10	8.68	8.68	2.71
Paramedic	0.92***	0.48°	8.76	8.38	1.79
CEO	1.49***	0.17	8.65	8.53	3.00
Investment Banker	0.96**	0.02	8.29	2.21	1.79

° $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$

Table S1. Item characteristics for each vignette.

Note. The first two columns give difference scores for each vignette, pooling data across studies, while the last three columns give the measurements of moral seriousness from Study S2. Significance levels are indicated for the Best–Middle and Middle–Worst comparisons.

Results and discussion. The vignette means for each measure are shown in Table S1, along with the mean differences for each comparison (Best–Middle, and Middle–Worst). The latter data were obtained by selecting all participants across Studies 1–5 in the main text, as well as supplementary Studies S3A, S3B, S4A, S4B, S5A, S5B, and S6 (the same studies used in the meta-analysis presented in the main text). The 7-point scales used in Study 1 were rescaled to a 9-point scale for consistency. The difference between Best and Middle was significant for all 8 vignettes. The difference between Middle and Worst reached marginal significance for only one out of the 7 vignettes, but was always numerically positive. This pattern is broadly consistent with the meta-analysis presented in the main text, which revealed robust differences between the Best and Middle conditions but only very small differences between the Middle and Worst conditions.

The Best–Middle difference — the measure of the optimality bias — was not significantly associated with severity [$r(6) = -.24$, $p = .56$], directness [$r(6) = -.19$, $p = .66$], or number [$r(6) = .28$, $p = .50$]. A multiple regression, adjusting for the effects of these variables simultaneously, reaches similar conclusions. Of course, strong conclusions about item differences cannot be drawn from a study of 8 vignettes.

The Middle–Worst difference also was not significantly associated with severity [$r(6) = .40$, $p = .33$], directness [$r(6) = .53$, $p = .17$], or number [$r(6) = –.35$, $p = .40$]. Although these correlations are not highly reliable, they are fairly large in magnitude. Perhaps when the harm is sufficiently serious, people are more inclined to attend not only to optimality, but also to more fine-grained probability information. This conclusion cannot be safely drawn from the current studies — particularly since the difference between Middle and Worst did not reach significance for most of the vignettes, taken individually — but is suggestive for future research.

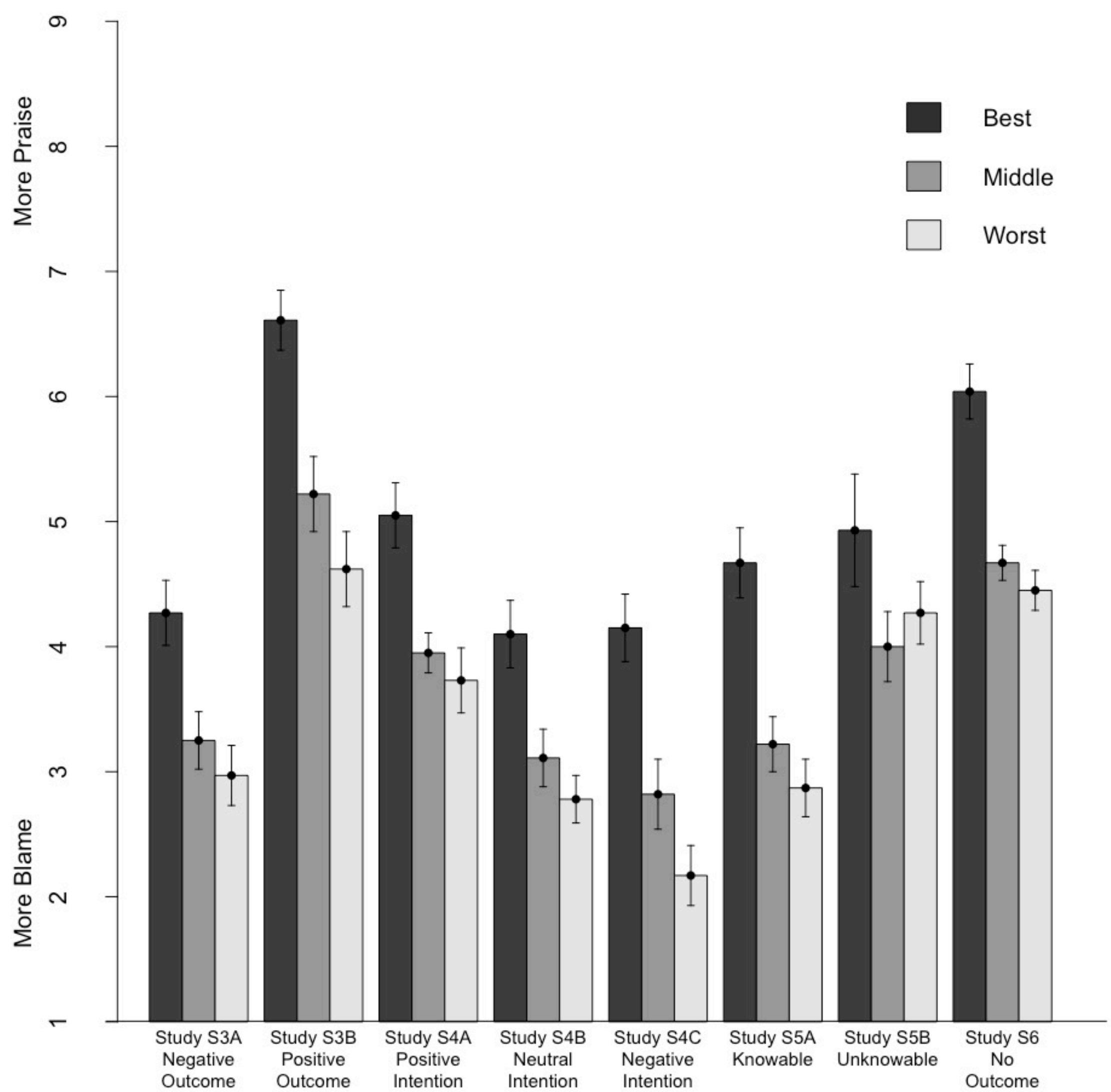

Figure S1. Results of Studies S3–S6.

Note. Bars represent 1 SE. Scales reverse-coded.

Study S3: Positive and Negative Outcomes

Although Study 4 demonstrated the optimality bias for positive outcomes, it did so in an unusual context where the outcomes are inevitable. Study S3 directly contrasted positive and negative outcomes in the more common situation where the outcomes are uncertain *ex ante*. In addition, this study allowed us to compare the magnitude of the *outcome bias* — the tendency to make harsher moral evaluations in light of negative outcomes (Baron & Hershey, 1988) — to that of the optimality bias.

Methods. We recruited 267 participants (M_{age} = 32, 39% female); 67 were excluded due to incorrect answers to check questions.

The method was the same as those in Study 1 in the main text (see Appendices S1–3), with three changes. First, the outcome was negative for some participants (Study S3A) and positive for others (Study S3B). Second, the information about the agent having done due diligence (used in Studies 2–7) was included. Finally, the dependent measure was moral blame (as in Studies 2–7).

Results and discussion. When judging the blameworthiness of the agents' actions for the negative outcomes in Study S3A, participants blamed agents choosing Best less than those choosing Middle [$t(64) = 2.86$, $p = .006$, $d = 0.71$, 95% CI[0.31,1.73], $BF_{10} = 6.5$, $d_S = 0.74$; Figure S1], but blamed agents choosing Middle no less than those choosing Worst [$t(71) = 0.82$, $p = .42$, $d = 0.19$, 95% CI[–0.40,0.95], $BF_{01} = 4.1$, $d_S = 0.67$]. Likewise, for the positive outcomes of Study S3B, moral judgments again tracked optimality, with more praise assigned to agents choosing Best than Middle [$t(61) = 3.57$, $p = .001$, $d = 0.90$, 95% CI[0.61,2.18], $BF_{10} = 40.2$, $d_S = 0.72$], but no more praise assigned to agents choosing Middle than Worst [$t(64) = 1.43$, $p = .16$, $d = 0.35$, 95% CI[–0.24,1.44], $BF_{01} = 2.1$, $d_S = 0.71$]. The interaction between experiment and condition was not significant [$F(2,194) = 0.85$, $p = .43$], failing to support a moderating effect of outcome. Hence, the same efficiency-based mechanisms appear to apply to moral judgments made in light of *both* negative and positive outcomes.

Although the patterns were similar for positive and negative outcomes, participants generally produced harsher evaluations of an agent's behavior in light of negative outcomes (Study S3A) than positive outcomes (Study S3B), even though the agent could control the outcome only indirectly through her choice [$t(198) = 8.47$, $p < .001$, $d = 1.20$, 95% CI[1.54,2.47], $BF_{10} > 1000$, $d_S = 0.39$], demonstrating a robust *outcome bias* (Baron & Hershey, 1988). Notably, the size of the optimality bias ($d = 0.71$ for negative outcomes and $d = 0.90$ for positive outcomes) approached the size of the outcome bias ($d = 1.20$).

Like Study 1, the omission of the unknowability stipulation from these stimuli (as well as those used in Study S4) raises important questions of internal validity: Could participants have believed that the agents did their research poorly, and hence blamed them for their ignorance? In the absence of strong stipulations about unknowability, this remains a possibility. Nonetheless, finding the same pattern as in the main experiments for both positive and negative outcomes suggests that the effect generalizes to cases where such stipulations are not made, which may better reflect the real world.

Study S4: Positive, Neutral, and Negative Intentions

One possible concern is that participants view the agents' choices as reflecting their *goals* — optimal decisions may have signaled a positive intention, whereas suboptimal decisions may not have. In that case, the increased blame for the suboptimal agents would reflect blame for the agent's mental states, not for their choice (although it is difficult to see, normatively, how the agents' positive or negative intentions could have driven which of the three options they chose, given their ignorance of the probabilities associated with these options). To address this possibility, Study S4A specified that the agent had a positive intention, and Study S4B specified that the agent did not have a positive intention (i.e., a "neutral" intention).

In addition, we examined a potential consequence of our efficiency account. From the moral *patient's* point of view, the agent's optimal choice is always Best. However, from the *agent's* perspective, the optimal choice depends on their intention. For a positively intentioned agent, the optimal choice is Best. But if the agent has negative intentions, the Worst option is actually optimal. Thus, perhaps when the agent has a negative intention, participants would distinguish between Middle and Worst — options that are always suboptimal from the patient's perspective, but which differ in optimality from the agent's perspective. Study S4C tested this possibility by specifying that the agent had a negative intention.

Methods. We recruited 503 participants (M_{age} = 34, 44% female); 202 were excluded because they incorrectly answered one or more check questions.

The method was the same as Study S3A, except that the agent's intent was specified to be either positive (e.g., "The doctor intends to choose the best treatment option for her patient"; Study S4A), neutral (e.g., "The doctor does not intend to choose the best treatment option for her patient"; Study S4B), or negative (e.g., "The doctor intends to choose the worst treatment option for her patient"; Study S4C). The outcome of the agent's action was always negative.

Results and discussion. Given positive intentions in Study S4A**,** judgments of blame for the positive outcome again tracked optimality, since agents were evaluated more positively if they chose Best than Middle [$t(76) = 3.54$, $p = .001$, $d = 0.80$, 95% CI[0.48,1.72], $BF_{10} = 39.5$, $d_S = 0.65$; Figure S1], but no differently if they chose Middle rather than Worst [$t(68) = 0.76$, $p = .45$, $d = 0.18$, 95% CI[–0.36,0.79], $BF_{01} = 4.2$, $d_S = 0.63$]. Similarly, given neutral intentions in Study S4B, judgments of blame tracked optimality, since agents were evaluated more positively if they chose Best than Middle [$t(65) = 2.64$, $p = .010$, $d = 0.65$, 95% CI[0.24,1.74], $BF_{10} = 4.0$, $d_S = 0.74$] but no differently if they chose Middle rather than Worst [$t(71) = 1.09$, $p = .28$, $d = 0.25$, 95% CI[–0.28,0.94], $BF_{01} = 3.3$, $d_S = 0.66$]. Thus, an optimal choice does not function merely as a signal of the agent's unrevealed intention, since the effect occurs even if the intention is specified. Comparing the means for Study S4A and S4B reveals a sizeable effect of positive versus neutral intention [$t(209) = 4.83$, $p < .001$, $d = 0.67$, 95% CI[0.59,1.40], $BF_{10} > 1000$, $d_S = 0.38$]. However, the effect size of the optimality bias was, if anything, even larger ($d = 0.80$ and $d = 0.65$) than the effect of intention ($d = 0.67$).

Unlike for positive and neutral intentions, for negative intentions there is a mismatch in optimality — Worst is the optimal choice from the agent's perspective, whereas Best is the optimal choice from the victim's perspective. Therefore, we predicted (*a priori*) that this clash in optimalities might cause some sign of a stepwise assignments of blame, unlike in the other conditions and the previous studies. Providing some evidence for this prediction, participants in Study S4C judged agents with negative intentions as more blameworthy when choosing Middle than Best [$t(53) = 2.84$, $p = .006$, $d = 0.76$, 95% CI[0.39,2.27], $BF_{10} = 6.2$, $d_S = 0.78$], but also marginally more blameworthy when choosing Worst than Middle [$t(61) = 1.78$, $p = .081$, $d = 0.45$, 95% CI[–0.08,1.38], $BF_{01} = 1.3$, $d_S = 0.76$]. However, the interaction between condition and experiment was not significant [$F(4,292) = 0.54$, $p = .70$], failing to support a moderating effect of intention.

One possibility is that the manipulation of intention was unsuccessful because participants drew exculpatory inferences about why the agents chose the worst option (e.g., maybe the patient desired that the doctor act the way she did), although this seems unlikely for many of the vignettes (e.g., it is not very plausible that the patient would ask for permanent hearing loss). Whatever the reason for these mixed results, they do not allow strong conclusions about intention-based differences in the optimality bias when the agent and patient have *different* optimal choices. Future research might further address the potential role of motivations on moral judgments under these particular circumstances.

Study S5: Knowable and Unknowable Probabilities

Studies 2–6 addressed the issue of perceived negligence (Nobes, Panagiotaki, & Pawson, 2009) in two ways — specifying in all cases that the agent did due diligence to assess the probabilities of the choices and that these probabilities were unknowable, and directly measuring attributions of negligence in Study 5. In Study S5, we further investigated the role of perceived negligence by comparing cases where the probabilities were unknown but *knowable* to cases where the probabilities were both unknown and *unknowable*.

Methods. We recruited 335 participants (M_{age} = 32, 37% female); 188 were excluded due to incorrect answers to check questions.

The vignettes were the same as those in Study S3A, except that participants were told either that evidence existed indicating that the options had different likelihoods of a positive outcome (*knowable* condition, e.g., "In fact, there is some existing evidence that says this belief is incorrect. As it happens, the doctor's belief is wrong."; Study S5A) or that no such evidence existed (*unknowable* condition, e.g., "In fact, all of the existing evidence says that this belief is correct. But as it happens, for reasons completely outside of her control, the doctor's belief is wrong."; Study S5B).

Results and discussion. In Study S5A, when the probabilities were knowable, participants judged the agent less harshly when she chose Best rather than Middle [$t(70) = 4.46, p < .001, d = 1.05$, 95% CI[0.80,2.09], $BF_{10} = 656.1$, $d_S = 0.67$; Figure S1], but equally harshly whether she chose Middle or Worst [$t(65) = 1.10, p = .27, d = 0.27$, 95% CI[–0.28,0.99], $BF_{01} = 3.1$, $d_S = 0.67$]. Likewise, in Study S5B, when the probabilities were unknowable, participants again judged the agent less harshly when she chose Best rather than Middle [$t(27) = 2.20, p = .037, d = 0.82$, 95% CI[0.06,1.80], $BF_{10} = 1.8$, $d_S = 1.10$], but equally harshly whether she chose Middle or Worst [$t(28) = –0.72, p = .48, d = –0.26$, 95% CI[–1.03,0.49], $BF_{01} = 3.1$, d_S = 1.06]. The interaction between condition and experiment was not significant [$F(2,141) = 2.02, p = .14$], failing to support a moderating effect of knowability. Thus, participants in the other studies do not seem to have been blaming suboptimal agents for their negligence in failing to know the probabilities, but rather for choosing suboptimally.

The effect of optimality was similar regardless of whether the outcome probabilities were knowable or unknowable. However, participants made harsher judgments overall when the probabilities were knowable (Study S5A) rather than unknowable (Study S5B) [$t(145) = 2.97, p = .003, d = 0.53$, 95% CI[0.26,1.27], BF_{10} = 8.3, $d_S = 0.39$]. This suggests that people expect moral agents to make exhaustive efforts to inform themselves about their decision, and will hold them accountable for not doing so. However, the size of this negligence effect (*d* = 0.53) was only about half that of the optimality effect (*d* = 1.05 for Study S5A and *d* = 0.82 for Study S5B).

Study S6: Unknown Outcomes

Study S6 tested two further questions about efficiency-based moral judgment. First, would this effect occur even when the outcome is unknown? Given that moral judgments are often tied to concrete outcomes (e.g., Baron & Hershey, 1988), people may only judge the agent's suboptimal choice relative to a particular outcome. This can lead to a biased search for justifications in light of the outcome (see Alicke, 2000). When the outcome is negative, people may search for reasons to justify negative judgments (and suboptimality would be one such reason) and when the outcome is positive, people may search for reasons to justify positive judgments (and optimality would again be one such reason). But when there is no outcome, this sort of outcome-based reasoning would be avoided altogether, which might erase the optimality bias (see Tversky & Shafir, 1992 for a similar phenomenon). Study S6 tested this possibility by omitting the outcome.

Second, to what extent do participants have conscious awareness of using efficiency-based thinking? Study 7 in the main text found that people mispredict their blame judgments for ignorant agents, believing that they would either differentiate among all three choices or give similar judgments to all three choices. This suggests that people lack awareness of using the Efficiency Principle. To further test this issue, participants in Study S6 were asked to choose a justification for their judgments.

Methods. We recruited 336 participants (M_{age} = 33, 45% female); 90 were excluded due to incorrect answers to check questions.

The method was the same as Study S5B, with two changes. First, the outcome was omitted from the vignettes. Second, on the page after the main measures, participants were asked to justify their responses by selecting one of three options — a mentalistic justification ("The doctor believed that she gave her patient the best treatment available, and couldn't have known otherwise, therefore she is not responsible for the harm which this choice may cause"), an efficiency-based justification (e.g., "The fact that the doctor didn't know which treatment was best does not remove her responsibility for the harm which this choice may cause"), or 'other' (writing in their own justification). The order of the mentalistic and efficiency-based justifications was randomized.

Results and discussion. Although Study S6 omitted outcomes, judgments were again more positive for agents choosing Best rather than Middle [$t(169) = 6.49$, $p < .001$, $d = 0.99$, 95% CI[0.95,1.79], $BF_{10} > 1000$, $d_S = 0.44$; Figure S1], but did not differ between Middle and Worst [$t(163) = 1.01$, $p = .32$, $d = 0.16$, 95% CI[–0.20,0.63], $BF_{01} = 5.0$, $d_S = 0.42$]. Thus, efficiency-based judgments are not tied to concrete outcomes, but occur even when the outcome is not specified. Indeed, the size of the effect was no smaller when outcomes were omitted ($d = 0.99$) as compared to previous studies (varying from $d = 0.30$ to $d = 1.11$), suggesting that efficiency-based thinking is tied entirely to the *optimality of the agent's decision* rather than to the outcome.

Justification	Proportion	Means		
		Best	Middle	Worst
Mentalistic	65.0%	6.27	5.28	5.13
Efficiency-Based	27.6%	5.06	3.43	3.29
Other	7.3%	6.30	4.50	3.50

Table S2. Response patterns in Study S6.

Note. Means are broken down according to the justifications chosen by participants. Means for the 'other' justifications are highly imprecise given the small cell sizes (18 participants total across three between-subjects conditions).

Most participants appear to be unaware of using an efficiency-based strategy, because 65.0% of participants selected the mentalistic justification (which argued that the agent should be exculpated), compared to 27.6% who selected the efficiency-based justification (which held the agent blameworthy) and 7.3% who selected 'other' (see Table S2). Further, responses were similar regardless of the explicit justification provided: Both the mentalistic and efficiency-based participants distinguished between Best and Middle [$t(111) = 4.64$, $p < .001$, $d = 0.87$, 95% CI[0.57,1.42], $BF_{10} > 1000$, $d_S = 0.54$ for mentalistic; $t(42) = 3.88$, $p < .001$, $d = 1.22$, 95% CI[0.78,2.48], $BF_{10} = 77.2$, $d_S = 1.02$ for efficiency-based], while neither group distinguished between Middle and Worst [$t(103) = 0.74$, $p = .46$, $d = 0.15$, 95% CI[–0.25,0.54], $BF_{01} = 5.1$, $d_S = 0.52$ and $t(50) = 0.45$, $p = .66$, $d = 0.13$, 95% CI[–0.47,0.75], $BF_{01} = 4.4$, $d_S = 0.76$, respectively]. Thus, even participants who self-reported an exculpatory mentalistic justification used efficiency-based reasoning.

These results suggest, together with Study 7, that participants lacked introspective access to the effects of the Efficiency Principle on their behavior. Further, people appear to construct post hoc justifications for their efficiency-based inferences, since nearly two-thirds of participants chose the mentalistic justification. Although it is possible that some participants thought that both justifications had merit, the key point is that regardless of which justification was deemed more important, participants had the same degree of optimality bias. This makes the bias all the more troubling in everyday moral judgment, since people may have difficulty identifying when they or others are falling prey to it.

Appendix S1: Vignette Wordings

Note. Exact text corresponds to Study 1 in the main text. Cross-reference with Appendices S2 and S3 to construct the vignettes used in the other studies.

Doctor A doctor working in a hospital has a patient who is having hearing problems. This patient has three, and only three, treatment options. The doctor believes that all treatment options have a 70% chance of giving the patient a full, successful recovery. But in fact the doctor's belief is wrong. Actually:

1) If she gives the patient treatment LPN, there is a 70% chance the patient will have a full recovery.
2) If she gives the patient treatment PTY, there is a 50% chance the patient will have a full recovery.
3) If she gives the patient treatment NRW, there is a 30% chance the patient will have a full recovery.

The doctor chooses treatment LPN, and the patient does not recover at all. The patient now has permanent hearing loss.

Farmer A local farmer needs to add insecticide to all her crops. The farmer has three, and only three, insecticide options. The farmer believes that all options have a 70% chance of solving her insect problem while also leaving the local river completely uncontaminated. But in fact the farmer's belief is wrong. Actually:

1) If she uses brand LPN, there is a 70% chance the river will remain uncontaminated.

2) If she uses brand PTY, there is a 50% chance the river will remain uncontaminated.
3) If she uses brand NRW, there is a 30% chance the river will remain uncontaminated.
The farmer chooses option [LPN/PTY/NRW], and the river becomes contaminated. A few locals drink the water and get extremely sick.

Contractor A building contractor is deciding what cement to use to build the foundation of a new house. The contractor has three, and only three, cement options. The contractor believes that all options have a 70% chance of withstanding an earthquake, if one occurs. But in fact the contractor is wrong. Actually:
1) If he uses brand LPN, there is a 70% chance the house will withstand an earthquake.
2) If he uses brand PTY, there is a 50% chance the house will withstand an earthquake.
3) If he uses brand NRW, there is a 30% chance the house will withstand an earthquake.
The contractor chooses brand [LPN/PTY/NRW], and one day a brief earthquake occurs and the house very quickly collapses. Two people who were in the house are killed.

Programmer A computer programmer is deciding what anti-virus software to use in order to protect top-secret government information. The programmer has three, and only three, software options. The programmer believes that all software options have a 70% chance of saving the government's information in the unlikely event that a virus penetrates the system. But in fact the programmer is wrong. Actually:
1) If she uses software LPN, there is a 70% chance that the software will be able to save the information.
2) If she uses software PTY, there is a 50% chance that the software will be able to save the information.
3) If she uses software NRW, there is a 30% chance that the software will be able to save the information.
The programmer chooses software [LPN/PTY/NRW], and 5 months later a virus penetrates the system and all the information is immediately lost.

Jet Pilot A jet pilot is deciding what type of missile to use for a combat mission that will take place at a very high altitude. The pilot has three, and only three, missile options. The pilot believes that all missile options have a 70% chance of successfully hitting their target at this altitude. But in fact the pilot is wrong. Actually:
1) If he uses missile LPN, there is a 70% chance the missile will successfully hit its target.
2) If he uses missile PTY, there is a 50% chance the missile will successfully hit its target.
3) If he uses missile NRW, there is a 30% chance the missile will successfully hit its target.
The pilot chooses missile [LPN/PTY/NRW], and during combat the missile completely misses its target and instead hits a jet from the pilot's own air force. The pilots of that jet are immediately killed.

Paramedic A paramedic needs to drive to a hospital in order to deliver a kidney to a patient who is having an emergency kidney transplant. The paramedic has three, and only three route options to get to the hospital. The paramedic believes that all the options will give him a 70% chance of arriving in time to save the patient's life. But in fact the paramedic is wrong. Actually:
1) If he takes route LPN, there is a 70% chance he will successfully deliver the kidney in time.
2) If he takes route PTY, there is a 50% chance he will successfully deliver the kidney in time.
3) If he takes route NRW, there is a 30% chance he will successfully deliver the kidney in time.
The paramedic chooses route [LPN/PTY/NRW], and arrives far too late. The patient dies.

CEO The CEO of a company is deciding what type of material to use in order to make a new biohazard suite. The CEO has three, and only three options. The CEO believes that all three options have a 70% chance of protecting suit users from high levels of radiation. But in fact the CEO is wrong. Actually:
1) If he chooses material LPN, there is a 70% chance that suit users will be protected from high levels of radiation.
2) If he chooses material PTY, there is a 50% chance that suit users will be protected from high levels of radiation.
3) If he chooses material NRW, there is a 30% chance that suit users will be protected from high levels of radiation.
The CEO chooses material [LPN/PTY/NRW], and 2 months later 2 suit users die of radiation poisoning.

Broker	An investment broker is deciding which investment option to choose for her client's life savings. The investment broker has three, and only three, options. The investment broker believes that all three options have a 70% chance of surviving an economic downturn. But in fact the investment broker is wrong. Actually: 1) If she chooses option LPN, there is a 70% chance her client's life savings will survive an economic downturn. 2) If she chooses option PTY, there is a 50% chance her client's life savings will survive an economic downturn. 3) If she chooses option NRW, there is a 30% chance her client's life savings will survive an economic downturn. The investment broker chooses option [LPN/PTY/NRW], and 3 months later there is a brief economic downturn and her client immediately loses her entire life's savings.

Appendix S2: Condition Wordings

Note. Cross-reference with Appendices S1 and S3 to construct the vignettes used in each study.

A doctor working in a hospital has a patient who is having hearing problems. This patient has three, and only three, treatment options.

Positive intention	The doctor intends to choose the best treatment option for her patient.
Neutral intention	The doctor does not intend to choose the best treatment option for her patient.
Negative Intention	The doctor intends to choose the worst treatment option for her patient.
Ignorant	The doctor believes that all treatment options have a 70% chance of giving the patient a full, successful recovery.
Due diligence	Based on many articles that the doctor has carefully read in respected medical journals, she truly believes that all three options have a 70% chance of giving the patient a full, successful recovery.
Knowledgeable	The doctor knows that:
Knowability unspecified	But in fact the doctor's belief is wrong. Actually:
Knowable	In fact, there is some existing evidence that says this belief is incorrect. As it happens, the doctor's belief is wrong. Actually:
Unknowable	In fact, all of the existing evidence says that this belief is correct. But as it happens, for reasons completely outside of her control, the doctor's belief is wrong. Actually:

1) If she gives the patient treatment LPN, there is a 70% chance the patient will have a full recovery.

2) If she gives the patient treatment PTY, there is a 50% chance the patient will have a full recovery.

3) If she gives the patient treatment NRW, there is a 30% chance the patient will have a full recovery.

Negative outcome	The doctor chooses treatment [LPN/PTY/NRW], and the patient does not recover at all. The patient now has permanent hearing loss.
Positive outcome	The doctor chooses treatment [LPN/PTY/NRW], and the patient

	recovers. The patient has no permanent hearing loss.
Unspecified outcome	The doctor chooses treatment [LPN/PTY/NRW].
Hypothetical negative outcome	Imagine that the doctor chose treatment [LPN/PTY/NRW], and that the patient did not recover at all, suffering permanent hearing loss.
Inevitable	We now also know that the patient had a gene that would have allowed any treatment to cure the disease.
Wrongness	How wrong was the doctor's behavior?
Punishment	How much should the doctor be punished?
Praise/blame	What does the doctor deserve to receive for her behavior?
Explanation	To what extent do you feel that an explanation is necessary for the doctor's choice?
Negligence	While answering the question about blame, did you think that if the doctor had thought more carefully or done more research, then she would have been able to know which options were better and which were worse?
Egocentrism	While answering the question about blame, did you think that the doctor had some sense of which options were better and which were worse?
Hypothetical praise/blame	What would the doctor deserve to receive for her behavior?

Appendix S3: Structure of Cases

Note. Cross-reference with Appendices S1 and S2 to construct the vignettes used in each study.

Study	Intent	Knowledge	Knowable	Outcome	Dependent Measures
1A	—	Ignorant	—	Negative	Wrongness
1B	—	Ignorant	—	Negative	Punishment
2	—	Due Diligence	Unknowable	Negative	Praise/Blame
3	—	Due Diligence	Unknowable	Negative	Explanation Praise/Blame
4	—	Due Diligence	Unknowable	Positive / Inevitable	Praise/Blame
5	—	Due Diligence	Unknowable	Negative	Praise/Blame Negligence Egocentrism
6A	—	Due Diligence	Unknowable	Negative	Wrongness Praise/Blame

6B	—	Knowledgeable	—	Negative	Wrongness Praise/Blame
7	—	Due Diligence	Unknowable	Hypothetical Negative	Hypothetical Praise/Blame
S3A	—	Due Diligence	—	Negative	Praise/Blame
S3B	—	Due Diligence	—	Positive	Praise/Blame
S4A	Positive	Due Diligence	—	Negative	Praise/Blame
S4B	Neutral	Due Diligence	—	Negative	Praise/Blame
S4C	Negative	Due Diligence	—	Negative	Praise/Blame
S5A	—	Due Diligence	Knowable	Negative	Praise/Blame
S5B	—	Due Diligence	Unknowable	Negative	Praise/Blame
S6	—	Due Diligence	Unknowable	Unspecified	Praise/Blame

Appendix S4: Sample Size Planning

Sample sizes for all experiments were planned *a priori*, using effect size estimates based on related studies (Johnson & Rips, 2014, *under review*). For most of the between-subjects studies (Studies 1, 3, 4 and S3–S5), we aimed for a sample size of 168 per study, in order to achieve 90% power. A somewhat smaller sample size was used in Study 3, due to a participant recruitment error.

We doubled the planned sample size (target N = 336, with small fluctuations across studies) for Studies 2, 5, and S6, for the following reasons. In the case of Study 2, a pilot study found a similar pattern of means but was underpowered due to a smaller effect size (compared to the other studies). In the case of Studies 5 and S6, we planned to study the relationship between participants' moral judgments and their responses to other questions (egocentrism and negligence attributions in Study 5 and justifications in Study S6), requiring a larger sample size.

We planned a smaller sample size (target N = 96, with small fluctuations across studies) in Studies 6 and 7, compared to the between-subjects studies. The studies manipulated the key independent variable within-subjects, so fewer participants were needed to achieve similar levels of power.

Sample sizes varied slightly from the target in some studies due to recruitment procedures. Sensitivity power analyses are reported for all *t*-tests, with the d_S statistics referring to the minimum effect size that would be detectable with 80% power given the actual sample size (after exclusions).

Appendix S5: Exclusion Criteria

Check Questions. Check questions were used in all studies in the main article and Supplementary Materials. For Studies 2–5, participants answered the following questions for their assigned vignette:

Belief. "TRUE or FALSE: The doctor believed that all [three treatments] had a 70% chance of leading to [recovery]."

Choice. "Given the treatment that the doctor chose, what was the actual chance of [that treatment leading to recovery]?"

Knowability. "Did the doctor have any way of knowing that this belief [about the probabilities] was false, or was it outside of her control?" (answer options: "Yes, there was evidence saying that her belief was incorrect" or "No, it was outside her control")

For Norming Studies S1 and S2, participants answered the following questions:

Choice. "What was the chance of success for the option which each person chose?"

Outcome. "Were the outcomes of these decisions generally good or bad?"

For Studies 1, 6, 7, and S3–S6, participants answered a subset of the following questions:

Belief. "Did the doctor know about the actual chances of success for each of the options?"

Choice. "What was the actual chance of success for the option which the doctor chose?"

Intention. "Did the doctor intend to choose the best treatment option for her patient?"

Knowability. "Was there existing evidence about the actual chances of success for each of the options?"

Specifically, participants in Studies 6 and 7 answered a generalized version of the belief question, a generalized version of the knowability question (Studies 6A and 7), and the choice question for each vignette (Studies 6A and 6B). Participants in Studies 1 and S3–S6 answered the belief and choice questions, as well as either the intention question (Study S4 only) or knowability question (Studies S5 and S6).

Exclusion criteria. Because our hypotheses are predicated on the assumption that participants understand the agent's belief and choice, as well as the unknowability of the probabilities, participants incorrectly answering any of these questions were excluded from analysis for most of the studies (Studies 1–4 and S3–S5) and for the norming studies (S1–S2). However, the results generally did not depend on this decision. When the analyses were repeated on all participants for these studies (1–4 and S3–S5), the comparisons had similar significance levels as the primary analyses in nearly all cases (see Table S3).

For the remaining studies (Studies 5–7 and S6), we planned to look at individual differences and to compare the results across studies. Many participants in these studies misunderstood the "knowability" question asked in some of these studies. Therefore, we excluded any participant from these studies who incorrectly answered any question *except* the knowability question. Once again, this decision did not affect the outcomes of the analyses. When the analyses were repeated on only participants answering all questions correctly, all Best–Middle and Middle–Worst comparisons had similar significance levels.

Study	Best–Middle (*p*)		Middle–Worst (*p*)	
	Main Analysis	Alternative Criteria	Main Analysis	Alternative Criteria
1A	<.001	<.001	.78	.85
1B	<.001	<.001	.11	.30
2	.036	.20	.25	.098
3	.027	.052	.79	.088
4	.001	<.001	.67	.76
5	<.001	.018	.77	.37
6A	<.001	<.001	.066	.44
6B	<.001	<.001	<.001	<.001
7	<.001	.023	<.001	.001
S3A	.006	.001	.42	.66

S3B	.001	.001	.16	.24
S4A	.001	.001	.45	.026
S4B	.010	.002	.28	.15
S4C	.006	.001	.081	.12
S5A	<.001	<.001	.27	.69
S5B	.037	<.001	.48	.76
S6	<.001	<.001	.32	.37

Table S3. Response patterns in Study S6.

Note. Entries are *p*-values for the primary dependent measure in each study (blame and wrongness are averaged for Study 6), for the comparisons between the Best and Middle conditions and between the Middle and Worst conditions. For each pair of columns, these analyses are reported first using the exclusion criteria for the main text analyses and second using the alternative exclusion criteria described above.