Cognition 214 (2021) 104770

Contents lists available at ScienceDirect

Cognition

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

“False positive” emotions, responsibility, and moral character

Rajen A. Anderson *, Rachana Kamtekar, Shaun Nichols, David A. Pizarro

Cornell University, United States of America

ARTICLE INFO

Keywords:
Moral judgment
Moral character
Guilt
Emotions

ABSTRACT

People often feel guilt for accidents—negative events that they did not intend or have any control over. Why might this be the case? Are there reputational benefits to doing so? Across six studies, we find support for the hypothesis that observers expect “false positive” emotions from agents during a moral encounter – emotions that are not normatively appropriate for the situation but still trigger in response to that situation. For example, if a person accidentally spills coffee on someone, most normative accounts of blame would hold that the person is not blameworthy, as the spill was accidental. Self-blame (and the guilt that accompanies it) would thus be an inappropriate response. However, in Studies 1–2 we find that observers rate an agent who feels guilt, compared to an agent who feels no guilt, as a better person, as less blameworthy for the accident, and as less likely to commit moral offenses. These attributions of moral character extend to other moral emotions like gratitude, but not to nonmoral emotions like fear, and are not driven by perceived differences in overall emotionality (Study 3). In Study 4, we demonstrate that agents who feel extremely high levels of inappropriate (false positive) guilt (e.g., agents who experience guilt but are not at all causally linked to the accident) are not perceived as having a better moral character, suggesting that merely feeling guilty is not sufficient to receive a boost in judgments of character. In Study 5, using a trust game design, we find that observers are more willing to trust others who experience false positive guilt compared to those who do not. In Study 6, we find that false positive experiences of guilt may actually be a reliable predictor of underlying moral character: self-reported predicted guilt in response to accidents negatively correlates with higher scores on a psychopathy scale.

1. Introduction

> “Everyone has told him and he knows there was nothing he could do and it’s not his fault, but he can’t sleep and he feels guilty about living life if she can’t. We were to go to the beach yesterday, but he didn’t go because he says if she can’t go to the beach why should I get to go.”
>
> —D., referring to her husband, who accidentally killed another person

The above quote comes from the website accidentalimpacts.org, an online community that provides support for people who have, accidentally and without any fault, caused severe injury or death to another person. The testimonials on the site chronicle the experience of many individuals who live with feelings of deep guilt over the consequences of their accidental actions. Indeed, those feelings of guilt appear to be so ubiquitous that there is a section of the website dedicated to helping people deal with the moral injury caused by their accidental actions.

At first glance, cases like these seem puzzling. If an action was truly accidental, an individual should neither receive blame nor blame themselves for that action.[1] There is a large body of work in the psychology of moral responsibility linking intentional action and the attribution of moral culpability (e.g., Cushman, 2008; Malle, Guglielmo, & Monroe, 2014; Shaver, 1985; Weiner, 1995), and an agent is more likely to be blamed when she intentionally brings about a harmful outcome (e.g., Cushman, 2008; Sloman, Fernbach, & Ewing, 2009). Accordingly, an agent who accidentally harms someone is likely to be judged as less blameworthy than an agent who intentionally harms someone (e.g., Armsby, 1971; Darley, Klosson, & Zanna, 1978; Darley & Shultz, 1990; Shultz, Wright, & Schleifer, 1986). These findings are consistent with normative accounts of moral blame or fault in philosophy and law that hold that an agent should only be blamed or faulted if the harm he caused was “in the sphere of the agent’s rational control” (Royzman & Kumar, 2004; cf. Badar & Marchuk, 2013; Fischer &

* Corresponding author at: Uris Hall, Cornell University, Ithaca, NY 14850, United States of America.
E-mail address: raa255@cornell.edu (R.A. Anderson).

[1] True accidents do not include unintended harmful outcomes that occur due to negligence and recklessness. People do assign blame for negligence and recklessness (e.g., Alicke, 1992; Raz, 2010; Sher, 2009).

https://doi.org/10.1016/j.cognition.2021.104770
Received 1 October 2020; Received in revised form 4 May 2021; Accepted 8 May 2021
0010-0277/© 2021 Elsevier B.V. All rights reserved.

Ravizza, 1998; Perkins, 1939; Vargas, 2013).

There is some evidence that similar attributional processes are at work when agents evaluate their own actions. Making an attribution that one is morally responsible – that one intentionally caused a harmful/immoral outcome – often results in a feeling of guilt, suggesting that the agent is assigning at least partial responsibility for the negative outcome to themselves (e.g., Hoffman, 1982; Mandel & Dhami, 2005; Smith, Webster, Parrott, & Eyre, 2002; Weiner, Graham, & Chandler, 1982). For example, Mandel and Dhami (2005) found that the amount of guilt experienced by prisoners convicted of various crimes was strongly associated with their amount of self-blame. In the absence of moral responsibility, however, theories of blame would predict that people should feel little guilt for committing a purely accidental harm.

However, as we described above, there are a great number of people who cannot seem to avoid feeling guilty even when they do not meet the criteria for moral responsibility. The philosopher Bernard Williams discusses cases like these in his essay *Moral Luck* (Williams, 1981). He asks his readers to imagine an accident in which a lorry driver, through no fault of his own, runs over and kills a child. Distraught over what has happened, the imagined lorry driver feels a great deal of guilt. As Williams points out, it would seem to an observer that the driver *should not* feel guilty: "Doubtless, and rightly, people will … try to move the driver from this state of feeling, move him indeed from where he is to something more like the place of a spectator". At the same time, Williams notes, observers would expect that the driver would need to be encouraged to take something more like a spectator's perspective on it, and "indeed some doubt would be felt about a driver who too blandly or readily moved to that position." (Williams, 1981 p.28). That is, while surely observers would try to dissuade the lorry from feeling this form of guilt for something that was not his fault, Williams believes that if the driver were persuaded *too* quickly, it would raise some eyebrows.

These cases of guilt for accidental actions highlight two puzzles (Kamtekar & Nichols, 2019). First, why do agents feel guilty for accidental harms when observers would not blame them to the same degree? Second, why do observers both 1) judge that such agents should receive less blame or feel less guilt and 2) disapprove if they do not at least initially feel some guilt?

1.1. *False positive emotions*

In the present paper, we aimed to examine this second puzzle by investigating the inferences that observers make of people who express (or fail to express) these "false positive" feelings (Sperber, 1996); that is, feelings that are not normatively appropriate but are nonetheless characteristically triggered by the situation. Feeling guilt for an accidental harm is a false positive response since you do not meet a necessary condition for guilt – that of being at fault. The distinction between false-positive and true positive emotions seems to apply to many kinds of emotions (see Kamtekar & Nichols, 2019 for discussion). Consider fear: if a person comes upon a rattlesnake on a trail, they will likely feel fear, and this is an appropriate or *true positive* instance of fear. The rattlesnake really does pose a danger. But people also often feel fear when they come upon a harmless garter snake. This would seem to be a false positive instance of fear, since the garter snake does not pose any danger.

One interesting question about false positive emotions is whether they are predictive of true positive emotions. If a person is not afraid of garter snakes does that mean they are likely to be unafraid of rattlesnakes? Will people rely on a person's false negative emotional responses to predict that person's true positive emotional responses? Our goal was to examine how people might use a specific person's display of a false positive moral emotions[2] (such as guilt for an accident, or gratitude toward a person who was simply performing a basic duty) – as a predictor of whether that person would feel "true positive" emotions (such as feeling guilty when they have actually committed an intentional harm). We also aim to examine whether false positive moral emotions predict something good about an agent's moral character and behavior. An important reason to investigate gratitude – true positive as well as false positive – alongside guilt is that gratitude is free of one potential confound one might worry about in the case of true positive versus false positive guilt. This is that true positive guilt requires the commission of a wrong, for which the agent may be faulted, and which would by itself result in a lowered assessment of the agent's character, with or without any information about their feelings of guilt. This is not the case for gratitude, since the subject feeling true or false positive gratitude is different from the agent who is going over and above their duty versus merely doing their duty.

1.2. *Inferring moral character*

We base our hypotheses on a growing body of literature that emphasizes the role of *character* in our moral judgments – people appear not just to evaluate the morality of particular *actions* but also the *agents* who commit those actions (for reviews, see Helzer & Critcher, 2018; Pizarro & Tannenbaum, 2012; Uhlmann, Pizarro, & Diermeier, 2015). Evaluations of moral character play an important role in how we think of other people: people prioritize moral character traits over other traits when judging the general positivity of a person (Goodwin, Piazza, & Rozin, 2014) and define personal identity largely in moral terms (Strohminger & Nichols, 2014). Furthermore, judgments of a person's morality more strongly predict liking and respect for that person than do judgments of that person's competence and sociability (Hartley et al., 2016).

When evaluating an agent's moral character, people are seeking to uncover the agent's "moral-cognitive machinery" (Helzer & Critcher, 2018) – the set of underlying psychological mechanisms that govern how that agent behaves regarding moral situations. People seek to infer the agent's intentions, motives, desires, meta-desires, beliefs, and other mental states (Ames & Johar, 2009; Critcher, Inbar, & Pizarro, 2013; Fedotova, Fincher, Goodwin, & Rozin, 2011; Gray, Young, & Waytz, 2012; Pizarro, Uhlmann, & Salovey, 2003;). From these psychological inferences, observers can then attempt to predict how that agent will behave in the future. This is consistent with what we know about the mechanisms underlying social prediction more generally, where individuals infer an agent's enduring traits and their temporary mental states from observable behavior, and then use those trait and state inferences to predict the agent's future behavior (Tamir & Thornton, 2018).

One specific method used to infer moral character is to attend to the emotions an agent displays regarding their moral behavior (Brandt & Reyna, 2011). Observers treat affective displays as potential sources of information about the agent's intentions and desires (Higgins, 1998). Whereas displays of positive affect might indicate that the agent is claiming ownership or responsibility of the action (e.g., Tracy & Robins, 2008; Weiner, 1985), negative affect might indicate that the agent is distancing themselves or repudiating the action (e.g., Gold & Weiner,

[2] For our purposes, "moral emotions" refers to emotions that are involved in facilitating prosocial behavior (e.g., compassion that motivates helping behavior), are responses to moral stimuli (e.g., anger at social injustice), or both (e.g., guilt for one's harmful actions that then leads to addressing those harms; Haidt, 2003). As two prototypical examples, in the present research we focus on guilt for one's own harms and gratitude for being the recipient of another's beneficence.

2000). For example, agents are judged more favorably when they perform prosocial behavior with a positive emotional display (e.g., smiling) or harmful behavior with a negative emotional display (e.g., grimacing) compared to when they perform those behaviors without the same emotional displays (Ames & Johar, 2009). This dynamic appears to play out in criminal courts – a defendant's perceived remorse is one of the most important factors in jurors' decisions of whether to give a death sentence (Haney, Sontag, & Constanzo, 1994).

1.3. The current studies

We hypothesized that even though blame and guilt are not normatively appropriate responses to having accidentally caused harm, an agent who *fails* to feel guilt for the accident will be considered atypical and judged as lacking in moral character, compared to an agent who does feel guilty for the accident. So, while it may be a normative error to feel guilt when one does not deserve blame, it is the sort of error that may benefit the agent because of what it communicates about their moral character.

In the current research, we investigated the relationship between expressions of false positive moral emotions (guilt and gratitude) and judgments of moral character (Studies 1–5) and the relationship between expression of false positive moral emotions and individual differences in moral traits (Study 6). Our main hypothesis was that observers would judge an agent who feels false positive moral emotions – one who feels guilt or gratitude in response to a situation that does not normatively warrant those emotions – to have a more positive moral character and to be more likely to feel those emotions in true positive cases than an agent who does not feel false positive moral emotions. See Table 1 for a summary of the studies and methods. All materials, data, analysis syntax, and preregistration information can be found on the Open Science Framework at https://osf.io/btwsq/. Per our preregistrations, analyses reported here exclude certain participants, although none of our conclusions are substantively altered if these participants are included (see OSF link).

2. Study 1

Our first study served as an initial test of our hypothesis, allowing us to examine the judgments that observers make of agents who feel the false positive moral emotions of guilt and gratitude. As our central focus, we wanted to test whether false positive moral emotions would be perceived as reliable predictors of an agent's moral character. We presented participants with two scenarios: one scenario involving an agent who felt guilt (or did not feel guilt) for an accident they caused, but for which they were not morally responsible, and one scenario involving an agent who felt gratitude (or did not feel gratitude) toward a service-worker who was merely doing their job (or toward a service-worker who acted above-and-beyond what their job required of them). We hypothesized that participants would have a more positive impression of the agent who felt guilt than of the agent who did not feel guilt, and that participants would have a more positive impression of the agent who felt gratitude toward someone who was just doing their job than of the agent who did not feel gratitude.

2.1. Method

2.1.1. Participants

We recruited 416 U.S. participants through the Amazon Mechanical Turk platform (*MTurk*), with the aim of recruiting at least 100 participants per condition, based on recommendations for achieving power > 0.80 for detecting moderate effect sizes (Brysbaert, 2019). Analyses were conducted only after all data were collected. Participants were excluded from analyses if they failed at least one of the two manipulation checks asking what happened in the vignettes ($N = 45$), leaving a final sample of 371 participants (54% female, $M_{age} = 38.87$).

Table 1
Overview of vignette designs and measures for Studies 1–6.

	Guilt vignette	Other emotion vignette	Measures
Study 1	Agent spills coffee on someone by accident. *Guilt* vs. *No Guilt*	Gratitude: Agent buys train ticket. 2 (time pressure: rush, no rush) X 2 (emotion: high gratitude, low gratitude)	Moral character; Social likability; Likelihood of future moral offense; Likelihood of future guilt and shame; Responsibility; Agent displayed right amount of emotion; Victim displayed right amount of emotion (only for Guilt)
Study 2	Agent spills coffee on someone by accident. 2 (responsibility: accident, reckless) X 2 (emotion: guilt, no guilt)	Gratitude: Agent buys train ticket. *Gratitude* vs. *No Gratitude*	Moral character; Social likability; Likelihood of future moral offense (only for Guilt); Likelihood of future charity (only for Gratitude); Likelihood of future guilt; Likelihood of future gratitude; Blame (only for Guilt); Praise (only for Gratitude); Agent felt right amount of emotion
Study 3	Agents accidentally lock out coworker. *Agent who experiences guilt* vs. *Agent who does not experience guilt*	Fear: Agents come across a harmless garter snake *Agent who experiences fear* vs. *Agent who does not experience fear*	Moral character; Social likability; Likelihood of future moral offense; Agent felt right amount of emotion; Blame (only for Guilt); How dangerous is a garter snake (only for Fear); likelihood of experiencing different emotions: happy, sad, anger, fear, guilt, pride, disgust
Study 4	Agent spills coffee on someone by accident *Guilt by agent* vs. *No guilt by agent* vs. *Vicarious guilt – near* vs. *Vicarious guilt – far*	None	Same as Study 2
Study 5	Agent spills coffee on someone by accident. *Agent who experiences guilt* vs. *Agent who does not experience guilt*	None	Using trust game design: choice between two potential interaction partners; money transferred to each partner; expected return from each partner
Study 6	Self-reported guilt for both unforeseen accident and foreseen but unintended harm	Gratitude: Self-reported gratitude for receiving both duty-driven help and exceptional help	Psychopathy; Machiavellianism; Narcissism; No Meaning in Life; Social Desirability

2.1.2. Design

All participants read two scenarios presented in random order and were asked the same series of questions regarding the individuals described in each scenario. In the *Coffee Spill* scenario, participants read about a woman (Janet) in a coffee shop who was walking toward the exit, failed to notice a wrapper on the floor, and slipped on it, spilling her drink on a man sitting nearby. The man, while annoyed, wiped his shirt off and told Janet "Hey, no worries. Accidents happen so don't feel bad." Participants then read one of two potential responses from Janet (between-subjects): Participants in the *guilt* condition read that Janet, with a guilty expression, told the man that he was right, but she still felt bad about it. Participants in the *no guilt* condition read that Janet, with a neutral expression, said to the man that he was right, so she did not feel bad about it.

Participants then rated Janet's moral character (how good a person Janet is and how much they would trust Janet) and Janet's social likability (how much they like Janet and how much they would want to get to know Janet; each on a scale from *1* = *not at all* to *7* = *a great deal*). They also made predictions of how much guilt they believed Janet would feel after having committed various moral infractions (stealing something from a store, rushing down the stairs and stepping on someone's foot; from *1* = *not at all* to *7* = *a great deal*), how much shame they believed Janet would feel if she stole something from a store (from *1* = *not at all* to *7* = *a great deal*), and how likely it was that Janet would commit a minor moral offense in the future (from *1* = *not at all* to *7* = *a great deal*). Participants also judged how responsible Janet was for what happened (from *1* = *not at all* to *7* = *a great deal*). Finally, participants rated whether Janet displayed the right amount of emotion, and whether the man who had coffee spilled on him displayed the right amount of emotion (from *1* = *she/he should have displayed much less emotion* to *7* = *she/he should have displayed much more emotion*).

In the *Train Ticket* scenario, participants read a short vignette about a man (Peter) in a train station who purchased a train ticket. Half of the participants were told that he was under time pressure to purchase the ticket (the train was leaving in fewer than 5 min), while the other half was told that he had plenty of time (the train was leaving in 30 min). In addition, half of the subjects were told that the man expressed a lot of gratitude toward the station agent for selling him the ticket and telling him how much time he had (i.e., "Wow, thank you SO much for your help!"), and the other half were told that he expressed low gratitude to the station agent (i.e., "Okay thanks"). The design was therefore a 2 (time pressure: rush, no rush) X 2 (emotion: high gratitude, low gratitude) between-subjects design. Participants then answered the same series of question (tailored to the *Train Ticket* scenario) as in the *Coffee Spill* scenario. After completing both scenarios, participants completed two attention checks, asking them to select what happened in each scenario.

2.2. Results and discussion

2.2.1. Coffee spill scenario

We combined the questions measuring participants' judgments of how good a person Janet is and how much they would trust Janet into a single index of moral character ($r_{Spearman\text{-}Brown} = 0.94$). We combined the questions measuring participant's judgment of how much they like Janet and how much they would want to get to know Janet into a single index of social likability ($r_{Spearman\text{-}Brown} = 0.94$). We also combined the three questions measuring Janet's predicted guilt from various moral infractions ($\alpha = 0.94$) into a single index of predicted guilt. For summary of results see Table 2.

Consistent with our hypotheses, participants rated Janet as having significantly better moral character, social likability, as being less likely to commit a minor moral offense, and as more likely to feel guilt and shame when she felt guilt than when she did not feel guilt, all $ps < 0.001$. Participants also judged that Janet should have displayed significantly more emotion in the *no guilt* condition than in the *guilt* condition, $p < .001$.

Together, these results suggest that people prefer agents who display guilt even for accidental acts. There were no significant differences in participants' judgments of whether Janet was responsible for the spill between the *guilt* and *no guilt* conditions, $p = .125$, or in judgments of whether the victim of the coffee spill displayed the right amount of emotion, $p = .30$.

Consistent with our primary hypothesis, participants treated the false positive expression of guilt (arising in response to an accident), as a positive predictor of an agent's moral character, and as predictive of how an agent would behave in cases where guilt would be normatively appropriate. Importantly, there were no differences in judgments of responsibility for the agent across conditions, despite participants reporting that in the *no guilt* condition Janet should have felt more guilt. Participants seemed to believe that Janet should feel some guilt, even if the harm was accidental.

Table 2
Results for the Guilt scenario (Study 1). Judgments of moral character served as our primary measure of interest, whereby agents who experience false positive guilt (vs. agents who experience no false positive guilt) are rated as having better moral character. This then has implications for the agent's judged social likability, likelihood of future moral offense, and likelihood of future guilt and shame for true positive situations.

	Guilt *M* (*SD*)	No guilt *M* (*SD*)	*t* (369)	*p*	*d*
Moral character	5.42 (1.03)	3.01 (1.32)	18.71	<0.001	1.94
Social likability	5.03 (1.18)	2.59 (1.50)	17.49	<0.001	1.81
Likelihood of future moral offense	2.99 (1.45)	4.81 (1.45)	12.10	<0.001	1.26
Likelihood of future guilt and shame	5.90 (1.13)	2.98 (1.59)	20.45	<0.001	2.13
Responsibility	4.08 (1.85)	4.36 (1.71)	1.54	0.125	0.16
Agent should have displayed more emotion	4.38 (0.73)	5.73 (1.47)	11.27	<0.001	1.17
Victim should have displayed more emotion	4.21 (0.78)	4.29 (0.70)	1.03	0.30	0.11

2.2.2. Train ticket scenario

As in the Coffee Spill scenario, we calculated a single index of moral character ($r_{Spearman\text{-}Brown} = 0.86$), social likability ($r_{Spearman\text{-}Brown} = 0.84$), and predicted guilt ($\alpha = 0.79$). Participants generally rated Peter more favorably when he displayed gratitude than when he did not display gratitude, and when he was not rushed than when he was rushed (see Fig. 1). "Grateful" Peter was rated as having better moral character, $F(1, 357) = 15.33$, $p < .001$, $\eta_p^2 = 0.04$, and greater social likeability, $F(1, 357) = 11.28$, $p = .001$, $\eta_p^2 = 0.03$. When Peter expressed high gratitude, he was also judged as less likely to commit a minor moral offense, $F(1, 357) = 5.17$, $p = .02$, $\eta_p^2 = 0.01$, as less responsible for what happened, $F(1, 357) = 6.08$, $p = .01$, $\eta_p^2 = 0.02$, as expressing more guilt for moral transgressions, $F(1, 357) = 23.68$, $p < .001$, $\eta_p^2 = 0.06$, and as not needing to display more gratitude, $F(1, 357) = 6.17$, $p = .01$, $\eta_p^2 = 0.02$.

Similarly, when Peter was not rushed, he was rated as having better moral character, $F(1, 357) = 19.28$, $p < .001$, $\eta_p^2 = 0.05$, as being more socially likable, $F(1, 357) = 7.41$, $p = .007$, $\eta_p^2 = 0.02$, as being less likely to commit a minor moral offense, $F(1, 357) = 5.98$, $p = .02$, $\eta_p^2 = 0.02$, as more responsible for what happened, $F(1, 357) = 6.08$, $p = .001$, $\eta_p^2 = 0.03$, as expressing more guilt for moral transgressions, $F(1, 357) = 3.97$, $p = .04$, $\eta_p^2 = 0.01$, and as not needing to display more gratitude, $F(1, 357) = 6.54$, $p = .01$, $\eta_p^2 = 0.02$. Contrary to our predictions, there were no significant interactions between expressions of gratitude and whether or not Peter was rushed, all $ps > 0.27$.

Participants made more positive judgments of Peter when he expressed gratitude than when he did not express gratitude, regardless of whether the gratitude was a true positive or false positive. One potential explanation for why we failed to find the predicted interactions is that the gratitude expressed by Peter never seemed excessive in the context of the scenario, and never appeared miscalibrated or "inappropriate." In addition, gratitude may be relatively less costly compared to guilt, in that guilt involves negative affect. There is therefore relatively little downside to feeling gratitude, even when it is unwarranted. Rather than seeing any of the conditions as "false positive" cases of gratitude, participants may have relied simply on whether Peter was grateful to the teller, and on whether he was conscientiousness enough to arrive at the train station on time.

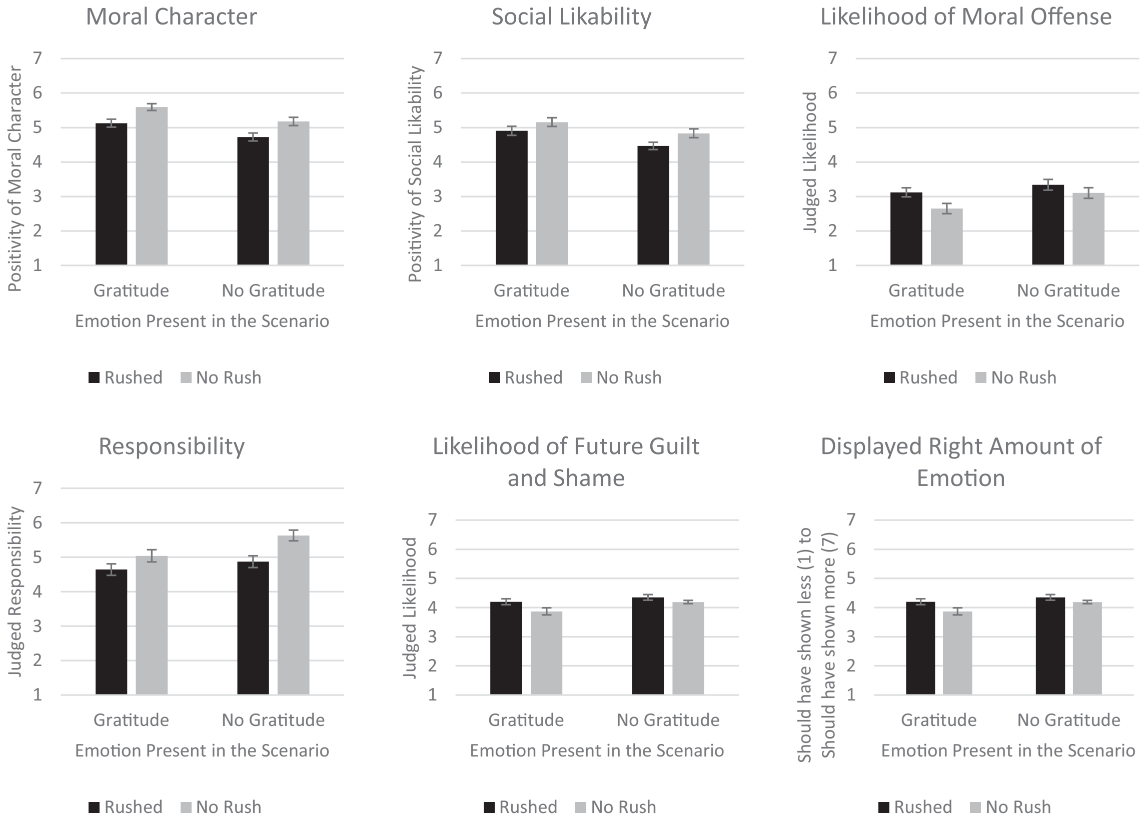

Fig. 1. Summary results for the Gratitude scenario (Study 1). Graphs display means and standard errors for each condition. The primary measure testing our hypothesis is the judgment of moral character.

3. Study 2

In Study 2 we aimed to both replicate and extend our findings from Study 1 by making several modifications to the materials and design. Specifically, we modified the guilt scenario to include a new set of conditions where the agent might be seen as having greater responsibility for the accident due to their own recklessness (i.e., having knowledge of the potential harmful consequences of an action and yet performing that action anyway). Varying whether an agent appears to have foreknowledge of the potential consequences of an action can influence the judgment of whether that action was done intentionally (Malle & Knobe, 1997; Perugini & Bagozzi, 2004). Observers may judge the agent as being more blameworthy, and their guilt as being more appropriate, for a case where the agent harms another person through recklessness) rather than as a completely unforeseen accident. We therefore explored the possibility that expressions of guilt may have a stronger effect on observers' impressions when an agent is harmed accidentally compared to when an agent is harmed due to recklessness.

We also included a modified version of the gratitude scenario from Study 1 in which an agent felt gratitude (or not) for someone else helping them while simply doing their job (this time without the time-pressure manipulation). Because gratitude is generally an emotion felt in response to another's moral or prosocial behavior toward the self (for a review, see McCullough, Kilpatrick, Emmons, & Larson, 2001), gratitude toward someone who is acting impersonally in order to fulfill their work duty might be viewed as a case of "false positive" gratitude. If an agent feels gratitude toward someone who assisted him solely because they are fulfilling their duty, observers may infer that the agent would feel grateful in a variety of other contexts and judge that agent as having good moral character.

3.1. Method

3.1.1. Participants

We recruited 408 U.S. participants through MTurk. Our initial aim was to recruit at least 100 participants per condition, which would provide power > 0.80 for our primary hypotheses based on the observed effect sizes in Study 1. We excluded participants if they failed the manipulation check (described below) for Scenario 1, leaving a final sample of 307 (56% female, $M_{age} = 38.97$). Contrary to our preregistration, we did not exclude participants for failing the manipulation check for Scenario 2 as all participants in the *no gratitude* condition failed the check. The relatively high failure rate for the Scenario 1 check (24.8%) and the very high failure rate for the Scenario 2 check (54.9%) suggests that the checks were overly difficulty for participants.

3.1.2. Design

Participants read two scenarios, presented in random order. In the *Coffee Spill* scenario, participants read an updated version of the *Coffee Spill* scenario from Study 1, based on a 2 (responsibility: accident, reckless) X 2 (emotion: guilt, no guilt) between-subjects design. As in Study 1, participants in the *accident* condition read that a woman, Janet, slipped on an empty wrapper on the floor and spilled her drink on a nearby man. Participants in the *reckless* condition read that Janet noticed a good friend outside the coffee shop and moved quickly to say hello, knowing that she might spill her drink, and then she bumped into an empty chair and tripped, spilling her drink on a nearby man. In both conditions, the man told Janet "Hey, no worries. Accidents happen so don't feel bad." To address concerns that the *publicly expressing* an emotion signals moral character (rather than simply *experiencing* an emotion), we changed the scenario such that the woman privately thought to herself either that she knew it was an accident but still felt bad about what happened (in the *guilt* condition), or that she knew it was an accident and did not feel bad about what happened (in the *no guilt* condition).

In the *Train Ticket* scenario, participants read an updated version of the scenario from Study 1, based on a 2-condition (emotion: gratitude, no gratitude) between-subjects design. As in Study 1, participants read about a man, Peter, going to a train station to buy a train ticket. Peter saw that the ticket counter was about to close for the day, so he rushed to the counter to buy his ticket. The ticket teller informed him that he arrived less than a minute before they were going to stop selling tickets. When Peter thanked the teller for staying open for him the teller responded, "Hey no worries, I'm just doing my job." We added this new statement from the teller to reinforce to participants that the teller believed he was merely doing his duty, to emphasize that gratitude for such behavior is not necessarily warranted. We then told participants that Peter either thought to himself "He was just doing his job, but I still feel grateful to him" (*gratitude* condition) or "He was just doing his job, so I don't actually feel grateful to him" (*no gratitude* condition).

Participants completed the same set of questions as in Study 1, presented in random order, for each scenario (unless otherwise noted, all items on a scale from *1 = not at all* to *7 = a great deal*). Participants were asked about the agent's moral character (how morally good Janet/Peter is, how good is Janet's/Peter's moral character, how much they would trust Janet/Peter), the agent's social likability (how much they like Janet/Peter and how much they would want to get to know Janet/Peter), were asked to predict how much guilt they believed Janet/Peter would feel after having committed various moral infractions (stealing something from a store, rushing down the stairs and stepping on someone's foot),[3] and were asked how much gratitude they believed Janet/Peter would feel after being the recipient of another's goodwill (a ticket teller having to stay open an extra couple minutes to serve her/him, a driver letting her/him ahead of him in traffic). To measure perceived moral culpability, we asked participants to assign blame for Janet/praise for Peter for what happened in each scenario. We also asked participants to predict the agent's future moral behavior (how likely it was that Janet would commit a minor moral offense, how likely it was that Peter would perform a small act of charity,). Finally, we asked participants whether Janet/Peter felt the right amount of either guilt (for Janet) or gratitude (for Peter) (from *1 = she/he should have felt much less guilt/gratitude* to *4 = she/he felt the right amount of guilt/gratitude* to *7 = she/he should have felt much more guilt/gratitude*).

Participants then completed a manipulation check for each scenario (see OSF link for details). The check for the *Coffee Spill* scenario asked participants what happened in the story with the woman at the coffee shop, and the check for the *Train Ticket* scenario asked participants how the man felt at the end of the train ticket story. Participants were considered to have passed the check if they selected the option that best summarized what happened in the scenario they read.

3.2. Results and discussion

3.2.1. Coffee spill scenario

We computed a composite index for moral character (how morally good Janet is, how good is Janet's moral character, how much they would trust Janet, $\alpha = 0.95$) and a composite index for social likability (how much they like Janet and how much they would want to get to know Janet, $r_{Spearman\text{-}Brown} = 0.93$). We also created composite indices of participants' predictions of Janet's guilt (guilt from stealing something from a store and from rushing down the stairs and stepping on someone's foot, $r_{Spearman\text{-}Brown} = 0.85$) and of Janet's gratitude (gratitude from a ticket teller having to stay open an extra couple minutes to serve her and from a driver letting her ahead of him in traffic, $r_{Spearman\text{-}Brown} = 0.92$). See Fig. 2 for a summary of the results.

Replicating our finding from Study 1 and consistent with our hypothesis, there was a significant main effect of emotion on judgments of general moral character, such that participants judged Janet as having better character when she felt guilt than when she felt no guilt, $F(1,303) = 231.45$, $p < .001$, $\eta_p^2 = 0.43$. There was no significant main effect of responsibility (i.e., accident vs. recklessness conditions) and no interaction between responsibility and guilt, $ps > 0.07$.

Likewise, participants judged Janet as being more socially likable when she felt guilt compared to when she felt no guilt, $F(1, 303) = 172.86$, $p < .001$, $\eta_p^2 = 0.36$. There was no significant main effect of responsibility (i.e., accident vs. recklessness conditions) and no interaction between responsibility and guilt, $ps > 0.25$.

Consistent with our hypotheses, participants judged Janet as more likely to commit a minor moral offense in the *no guilt* condition than in the *guilt* condition, $F(1, 303) = 97.96$, $p < .001$, $\eta_p^2 = 0.24$. There was no significant main effect of the responsibility manipulation, and no significant interaction between responsibility and guilt, $ps > 0.19$. Participants also judged Janet as being more likely to feel guilty in other situations in the *guilt* condition compared to the *no guilt* condition, $F(1, 303) = 372.43$, $p < .001$, $\eta_p^2 = 0.55$. For predictions of guilt, there was no significant main effect of responsibility and no interaction between the responsibility and guilt conditions, $ps > 0.68$.

Participants judged that Janet was more likely to feel gratitude in other situations in the *guilt* condition than in the *no guilt* condition, $F(1, 303) = 269.94$, $p < .001$, $\eta_p^2 = 0.47$. There was no significant main effect of the responsibility condition on such judgments, $F(1, 303) = 0.98$, $p = .32$, $\eta_p^2 = 0.003$, and no significant emotion by responsibility interaction, $F(1, 303) = 3.70$, $p = .055$, $\eta_p^2 = 0.01$.

Consistent with our predictions, for judgments of blame, there was a significant main effect of emotion such that participants judged Janet as more blameworthy in the *no guilt* condition than the *guilt* condition, $F(1,302) = 17.79$, $p < .001$, $\eta_p^2 = 0.06$. There was also a significant main effect of responsibility such that participants judged Janet as more blameworthy in the *reckless* condition than in the *accident* condition, $F(1, 302) = 111.93$, $p < .001$, $\eta_p^2 = 0.27$. These main effects were qualified by a significant interaction between emotion and responsibility, $F(1, 302) = 8.02$, $p = .005$, $\eta_p^2 = 0.03$. Breaking down this interaction, participants judged Janet as more blameworthy in the *no guilt* condition than in the *guilt* condition in the *accident* condition, $t(302) = 5.03$, $p < .001$, $d = 0.58$, but there was no significant difference in blame between the *no guilt* and *guilt* conditions in the *reckless* condition, $t(302) = 0.97$, $p = .33$, $d = 0.11$.

Confirming that the manipulation was effective, participants felt that Janet should have felt more guilt in the *no guilt* condition than in the *guilt* condition, $F(1, 302) = 34.79$, $p < .001$, $\eta_p^2 = 0.10$. There was no significant effect of responsibility on judgments of how much guilt Janet should have felt, $F(1, 302) = 3.369$, $p = .066$, $\eta_p^2 = 0.01$. There was no significant interaction between emotion and responsibility, $F(1, 302) = 0.49$, $p = .48$, $\eta_p^2 = 0.002$.

In summary, our results from the *Coffee Spill* scenario provide additional evidence that observers infer moral character from an agent's false positive expressions of guilt, and that they use these expressions to predict the agent's social likability and future moral behavior and reactions. There were more mixed effects with the responsibility manipulation – except for judgments of blameworthiness, participants were not sensitive to whether the behavior was accidental or due to recklessness. Instead, people appeared to be focusing primarily on the presence or absence of guilt in these vignettes. Interestingly, the presence or absence of guilt experienced by the agent in the *accident* conditions influenced how blameworthy the agent was judged by participants. One potential explanation for this effect is that participants interpreted the agent's own guilt as a form of self-blame, so when the agent did not feel guilty then participants increased their blame to account for the agent's lack of self-blame.

[3] Unlike in Study 1, we did not include any measure regarding the agent's tendency to experience shame.

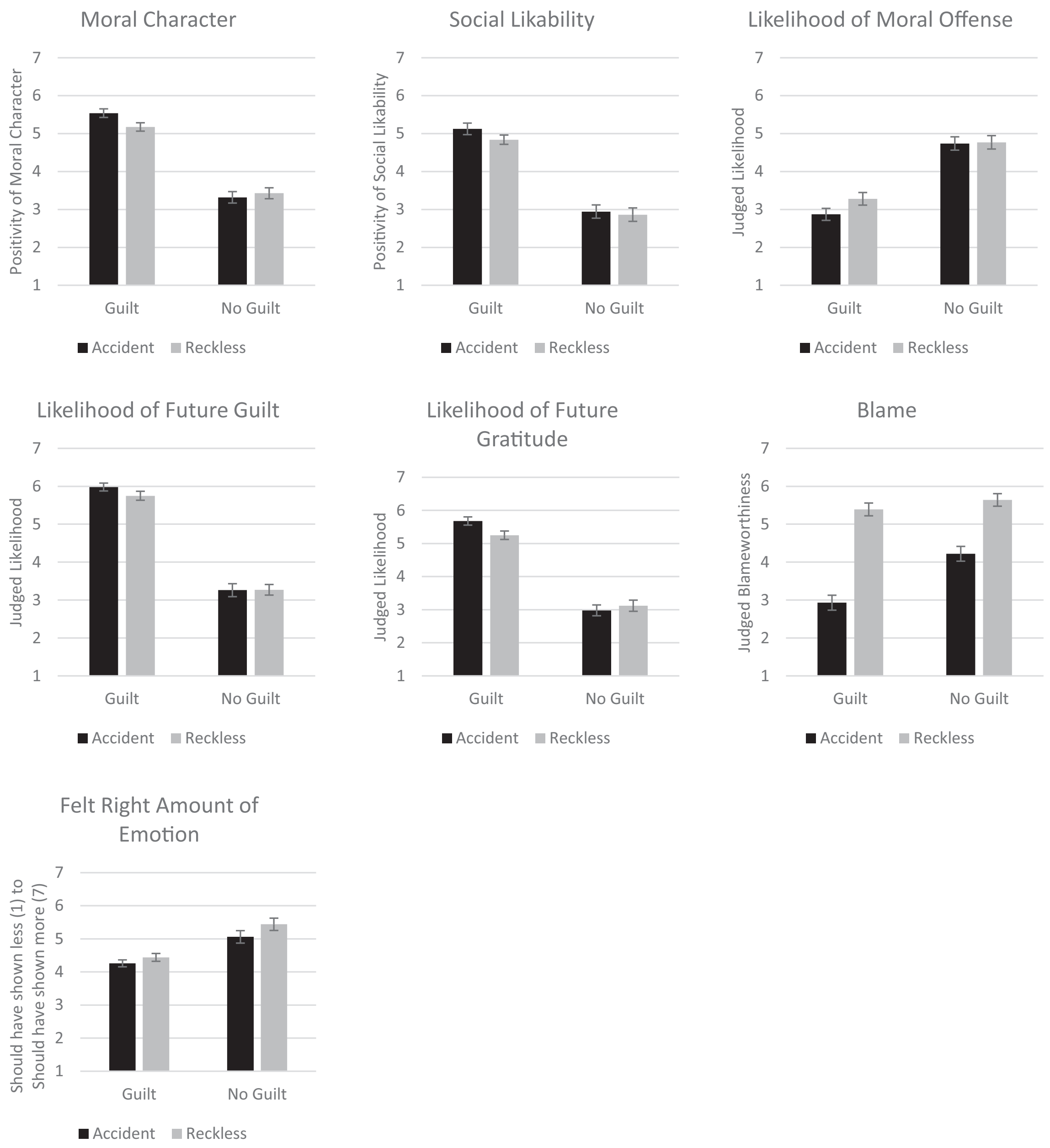

Fig. 2. Summary results for the Guilt scenario (Study 2). Graphs display means and standard errors for each condition. Judgments of moral character are the primary measure of interest.

3.2.2. *Train ticket scenario*

As with the *Coffee Spill* scenario, we combined items to form single measures of general moral character (α = 0.94), social likability ($r_{Spearman\text{-}Brown}$ = 0.91), predicted guilt ($r_{Spearman\text{-}Brown}$ = 0.85), and predicted gratitude ($r_{Spearman\text{-}Brown}$ = 0.89). See Table 3 for a summary of results. Consistent with our predictions, participants rated Peter as having better moral character in the *gratitude* condition than in the *no gratitude* condition, $p < .001$, and being more socially likable, $p < .001$. Furthermore, participants in the *gratitude* condition, relative to the *no gratitude* condition, rated the man as experiencing more guilt from moral infractions, $p < .001$, and more gratitude from others' kindness, $p < .001$. Participants rated Peter as more praiseworthy in the *gratitude* condition than the *no gratitude* condition, $p < .001$, and judged him as more likely to do a small act of charity, $p < .001$. Finally, participants reported that Peter should have felt more gratitude in the *no gratitude* condition than in the *gratitude* condition, $p = .02$. Much like guilt, false positive expressions of gratitude are treated by observers as predictors of

Table 3
Results for the Gratitude scenario (Study 2). Judgments of moral character served as our primary measure of interest.

	Gratitude *M* (*SD*)	No Gratitude *M* (*SD*)	*t* (305)	*p*	*d*
Moral character	5.67 (0.85)	3.90 (1.29)	14.19	<0.001	1.62
Social likability	5.55 (0.93)	3.46 (1.38)	15.84	<0.001	1.77
Likelihood of future guilt	5.99 (1.00)	4.25 (1.59)	11.47	<0.001	1.31
Likelihood of future gratitude	6.18 (0.82)	4.03 (1.69)	14.09	<0.001	1.62
Praise	4.25 (1.79)	2.55 (1.61)	8.76	<0.001	1.00
Likelihood of future act of charity	5.93 (1.01)	3.79 (1.48)	14.68	<0.001	1.68
Agent should have felt more emotion	4.28 (0.91)	4.60 (1.39)	2.36	0.02	0.27

an agent's moral character and future behavior. Even if gratitude is directed toward someone fulfilling their duties, observers treat such gratitude as indicative of the agent's character.

4. Study 3

In Study 3, we expanded our investigation connecting false positive emotions and judgments of moral character by including assessments of a wider array of emotions, in order to assess whether false positive expressions of *nonmoral* emotions would also be treated as predictors of a person's moral character. For example, if an agent were to feel fear at a harmless stimulus (i.e., a target that should not trigger fear), would observers infer that the agent has good moral character and would feel guilty for harm they have caused? One possibility is that observers would infer that a person who expresses false positive emotions of any kind (i.e., an emotional person) would be likely to express moral emotions in the future. However, we predicted that the expression of moral emotions like guilt would be especially tied to assessments of moral character, which themselves are fundamentally about an agent's underlying cognitive processes regarding moral decisions (e.g., Helzer & Critcher, 2018; Pizarro & Tannenbaum, 2012; Uhlmann et al., 2015). We reasoned that because of this, morally relevant expressions like guilt should be treated as more informative of a person's moral character than morally irrelevant expressions like fear. Specifically, we predicted that judgments of moral character would vary based on whether an agent felt guilty or not and would not vary based on whether an agent felt fear or not. That is, observers would treat different emotions as predicting different parts of an agent's underlying character. Finally, in Study 3 we also aimed to replicate the findings of Studies 1–2 regarding false positive expressions of guilt using a new scenario to ensure that the previous results were not simply artifacts of the particular stimuli we used (see Westfall, Judd, & Kenny, 2015).

4.1. Method

4.1.1. Participants

We recruited 120 U.S. participants (47% female, $M_{age} = 37.36$) through MTurk. This sample provides power > 0.95 to detect sample sizes observed in Studies 2. We did not include any comprehension checks or exclusion criteria.

4.1.2. Design

In a 2 (emotion type: guilt, fear) X 2 (emotion presence: agent felt the emotion, agent did not feel the emotion) within-subjects design, participants read two scenarios, presented in random order, and made judgments about two of the characters in each story. The names of the characters and the order in which participants made judgments of them was counterbalanced between participants.

In the *Guilt* scenario, participants read about two coworkers who were the last in the office to leave for lunch and, following standard practice at their work, locked the doors as they were the last to leave. When the coworkers returned from lunch, they saw a visiting colleague standing outside the door, who explained that he had been locked out for 45 min after returning from lunch because he did not know about the policy of locking the doors during lunch, and that he understood that it was a mistake that he was left waiting. After hearing this, one coworker felt guilty and thought to herself "I know it was just a misunderstanding, and we were following office policy, but I still feel bad that he was waiting so long.", while the other coworker did not feel guilty and thought to herself "I know it was just a misunderstanding, and we were following office policy, so I don't feel bad that he was waiting so long."

In the *Fear* scenario, participants read about two different coworkers returning to their workplace from lunch, taking the quickest path through a wooded park. As they walked through the park, they saw a large garter snake in the middle of the path, which looked at them a moment before moving off the path into the bushes. Upon first seeing the garter snake, one coworker felt afraid and thought to herself "I know it's just a harmless garter snake, but it still scares me a little.", while the other coworker did not feel afraid and thought to herself "I know it's just a harmless garter snake, so I don't feel scared at all."

For all four agents across the two scenarios, participants answered the same moral character items (all αs > 0.82) and social likability items (all $r_{Spearman\text{-}Brown}$s ≥ 0.82) from Study 2, the likelihood of committing a minor moral offense item from Study 2, and whether the agent felt the right amount of the emotion item adapted from Study 2. For the agents in the *Guilt* scenario, participants also reported how blameworthy each agent was for what happened (from *1* = *not at all* to *7* = *a great deal*). For agents in the *Fear* scenario, participants also reported how dangerous a garter snake is (from *1* = *harmless* to *7* = *extremely dangerous*). In addition, for all agents participants answered how likely each agent was to feel certain emotions in everyday life (guilt, anger, fear, sadness, happiness, disgust, and pride, from *1* = *not at all* to *7* = *a great deal*).

4.2. Results and discussion

Consistent with our hypothesis that participants use the presence of certain emotions to inform their judgments of moral character, we found a significant main effect of emotion presence, $F(1, 119) = 124.34$, $p < .001$, $\eta_p^2 = 0.51$, and the predicted interaction between emotion type and emotion presence on judgments of general moral character, $F(1, 119) = 99.52$, $p < .001$, $\eta_p^2 = 0.46$. Specifically, in the *Guilt* scenario participants rated the agent who *felt the emotion* as having better general moral character than agent who *did not feel the emotion*, but in the *Fear* scenario there was no such difference between the agent who *felt the emotion* and the agent who *did not feel the emotion* in general moral character (Fig. 3). There was a nonsignificant effect of the emotion type on judgments of moral character, $F(1, 119) = 0.63$, $p = .43$, $\eta_p^2 = 0.005$.

We next examined how participants assessed the agent's social likability. There was a significant main effect of emotion type, $F(1, 119) = 13.84$, $p < .001$, $\eta_p^2 = 0.10$, a significant main effect of emotion presence, $F(1, 119) = 51.65$, $p < .001$, $\eta_p^2 = 0.30$, and a significant interaction, $F(1, 119) = 73.17$, $p < .001$, $\eta_p^2 = 0.38$. Specifically, in the *Guilt* scenario participants rated the agent who *felt the emotion* as being more socially likable than agent who *did not feel the emotion*, but in the *Fear* scenario participants rated the agent who *felt the emotion* as less socially likeable than the agent who *did not feel the emotion*.

There was also a significant main effect of emotion presence, $F(1, 115) = 41.86$, $p < .001$, $\eta_p^2 = 0.27$, and a significant interaction between emotion type and emotion presence on the predicted likelihood of the agent committing a minor moral offense, $F(1, 115) = 11.01$, $p = .001$, $\eta_p^2 = 0.09$. Specifically, within each scenario, participants rated the agent who *felt the emotion* as being less likely to commit a minor moral offense than the agent who *did not feel the emotion*, but this difference was larger for agents within the *Guilt* scenario than for agents within the *Fear* scenario.

For evaluations of whether the agents felt the right amount of emotion, there was a significant main effect of both emotion type and emotion presence. Participants thought that agents in the fear scenario should have felt relatively less emotion than agents in the guilt scenario, $F(1, 118) = 9.59$, $p = .002$, $\eta_p^2 = 0.08$. In addition, participants thought that agents who *felt the emotion* should have felt relatively less of that emotion, while they thought agents who *did not feel the emotion* should have felt relatively more, $F(1, 118) = 59.32$, $p < .001$, $\eta_p^2 = 0.33$. There was no significant interaction between emotion type and presence, $F(1, 118) = 3.58$, $p = .06$, $\eta_p^2 = 0.03$. For the guilt scenario, there was no significant difference in blame between the agent who *felt the emotion* ($M = 2.15$, SD = 1.47) and the agent who *did not feel the emotion* ($M = 2.34$, SD = 1.52), $t(118) = 1.68$, $p = .096$, $d = 0.13$. This makes sense, as both agents were involved in the accident.

As predicted, for the *fear* scenario, there was no significant difference in the judged dangerousness of a garter snake when participants were answering in reference to the agent who *felt the emotion* ($M = 1.71$, SD =

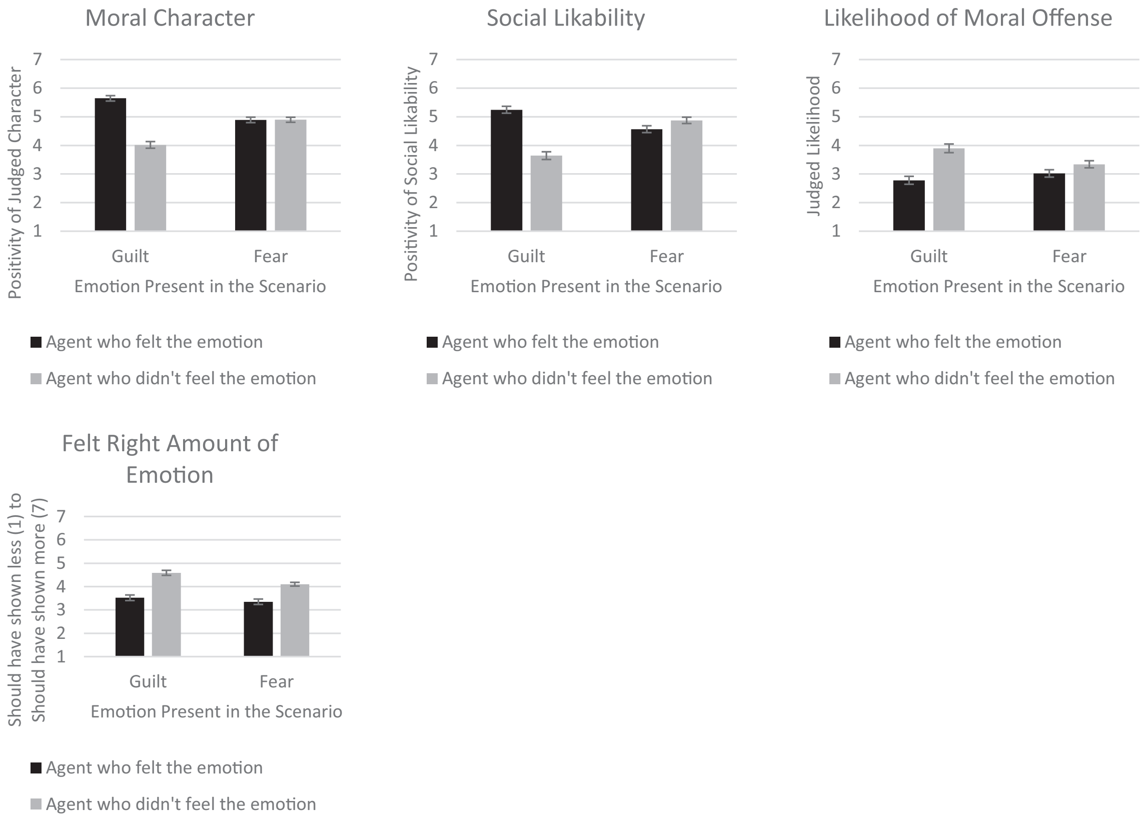

Fig. 3. Summary results (Study 3). Graphs display means and standard errors for each condition. Judgments of moral character served as the primary test of our hypothesis – it is not simply any false positive experience of emotion that observers use to infer character.

1.29) and the agent who *did not feel the emotion* ($M = 1.63$, SD = 1.17), $t(119) = 1.15$, $p = .25$, $d = 0.06$.

To further test the role of different emotions in judgments of moral character, we examined participants' predictions of the likelihood that different agents would experience various emotions (please see Supplementary Fig. 1). There was no clear pattern of effects of emotion type and emotion presence on these likelihood judgments. For example, participants predicted that the presence of an emotion in an agent, either guilt or fear, made it more likely that the agent would experience sadness (compared to no emotion), but made the opposite prediction regarding the likelihood of experiencing pride. Together, these inconsistent patterns of results suggest that participants are not using an individual's generalized tendency to experience emotions or an individual's overall level of emotionality, but are instead making more specific and nuanced inferences about the agents' emotionality and moral character based on the presence (or absence) of specific emotions in a contextually relevant scenario.

5. Study 4

In Study 4, we investigated whether our previous findings showing that expressions of guilt influence character evaluations of an agent were a result of the agent having been described as expressing *any* guilt at all. That is, people may form a positive impression of anyone who feels guilty concerning a harmful outcome. In Studies 1–3, the *guilt* and *no guilt* conditions differ both in whether the agent expressed guilt for an accidental harm and also whether the agent expresses *any* guilt at all. Therefore, it remained unclear whether the differences we observed were driven by the *false positive* guilt (as we would predict) or by the presence of guilt in general. If an agent felt guilty for an accidental harm that was entirely outside their causal control, would observers make the same inferences regarding their moral character as they would for an agent who felt guilty for an accidental harm they caused? Instead of being treated as a predictor of moral character, cases of extreme false positive guilt may actually be treated by observers as predictors of neuroticism or a pathological sense of responsibility.

We hypothesized that these expressions of "false positive" guilt would be informative about an agent's character when the agent has a *reasonable* counterfactual about how their causal role could have been different. Guilt is often a result of counterfactual thinking about an event (e.g., *what if I had done this instead?*), along with mental attempts to undo the harmful event (e.g., Davis, Lehman, Wortman, Silver, & Thompson, 1995; Kamtekar & Nichols, 2019; Mandel & Dhami, 2005; Niedenthal, Tangney, & Gavanski, 1994). However, it is likely that certain counterfactual thoughts felt by agents are too far-fetched and removed from the situation for the guilt to indicate that the agent can be reliably trusted and possesses good moral character. For example, in the Williams (1981) case of the lorry driver accidentally killing someone, if a friend of the driver and expressed guilt, exclaiming "If only I had called him and told him not to work today, this could have been prevented!", observers might feel that such guilt was excessive, if not overly dramatic. Guilt over such an unrealistic counterfactual would seem to say little about the friend's moral character (but would perhaps be informative about his other psychological qualities). Accordingly, we predicted that agents who expressed guilt for harmful accidents that were entirely outside of their causal control would not be evaluated as morally positively as agents who expressed guilt for harmful accidents in which they played a causal (but accidental) role.

5.1. Methods

5.1.1. Participants

We recruited 438 U.S. participants through MTurk. We aimed for at least 100 participants per condition to achieve power > 0.90 based on the observed effects in Studies 1–3. Per our preregistration, we excluded 3 participants for completing the study in less than 30 s, leaving us with a final sample size of 435 (46% female, $M_{age} = 36.77$).

5.1.2. Design

We randomly assigned participants to one of four conditions. All participants read a modified version of the *accident* version of the *Coffee Spill* scenario from Study 2, in which a woman at a coffee shop, Janet, accidentally spills her drink on another customer. Participants in the *guilt* condition read that Janet thought to herself “It’s too bad that his shirt is stained, and even though it was an accident, I still feel guilty about it”, then apologized to the man and helped him clean up. Participants in the *no guilt* condition read that Janet thought to herself “It’s too bad that his shirt is stained, but it was an accident, so I don’t feel guilty about it”, then apologized to the man and helped him clean up. Additionally, we included two conditions in which participants read about Janet spilling her drink, apologizing to the man, and helping him clean up, but with no mention of her own feelings. However, she then relates the events to a friend (Sarah). Participants in the *vicarious guilt – near* condition read that Janet was at the coffee shop to meet her friend Sarah, who had arrived at the coffee shop at the agreed-upon time—-right after the accident. Janet then told Sarah about the accident, and Sarah thought to herself “It’s too bad that I wasn’t there when it happened. I know I arrived on time, but if only I had gotten here a little earlier, I would have been able to prevent this from happening. I feel guilty that I wasn’t able to stop this.” Participants in the *vicarious guilt – far* condition read that later in the day Janet phoned Sarah and told her about the accident at the coffee shop. Sarah then thought to herself “It’s too bad that I wasn’t there when it happened. I know I don’t live there, but if only I was visiting her at the time, I would have been able to prevent this from happening. I feel guilty that I wasn’t able to stop this.”

We then asked participants to complete the same measures from Study 2 (moral character [α = 0.88], social likability [$r_{Spearman\text{-}Brown} = 0.87$], predicted guilt [$r_{Spearman\text{-}Brown} = 0.75$], predicted gratitude [$r_{Spearman\text{-}Brown} = 0.77$], blame, agent’s future moral behavior, and feeling the right amount of guilt), with some modifications. Participants who read the Janet *guilt* or the Janet *no guilt* scenario responded to the questions as pertaining to Janet, whereas participants who read the *vicarious guilt-near* or the *vicarious guilt-far* scenario answered the questions as pertaining to Sarah. In addition, we added two items assessing the neuroticism of the agent (i.e., (i) whether they would describe Janet/Sarah as someone who worries a lot, and (ii) whether they would describe Janet/Sarah as someone who is emotionally stable and as someone who is not easily upset (reverse-coded), $r_{Spearman\text{-}Brown} = 0.61$, from *1 = not at all* to *7 = very much*).

5.2. Results and discussion

Per our preregistration, we conducted an omnibus ANOVA for each of the measures, followed-up with planned contrasts comparing responses from participants in the Janet *guilt* condition to responses from participants in each of the other three conditions (Janet *no-guilt*, *vicarious guilt-near*, and *vicarious guilt-far*). See Table 4 for descriptive statistics.

Consistent with our hypothesis, there was a significant difference between conditions on judgments of moral character, $F(3, 431) = 24.67$, $p < .001$, $\eta_p^2 = 0.15$. Planned contrasts revealed a significant difference in judgments of moral character between the Janet *guilt* condition and the Janet *no guilt condition*, $t(431) = 7.34$, $p < .001$, $d = 1.03$; a significant difference between the Janet *guilt* condition and the *vicarious guilt – far* condition, $t(431) = 3.02$, $p = .003$, $d = 0.39$; and no significant difference between the Janet *guilt* condition and the *vicarious guilt – near* condition, $t(431) = 0.15$, $p = .88$, $d = 0.02$. While the lack of a significant difference between the *guilt* and *vicarious guilt – near* conditions was unexpected, it is possible that participants viewed Sarah’s guilt in the latter condition as a sign of empathy and a recognition that she actually could have helped had she been there slightly earlier. In other words, perhaps our participants did not view Sarah’s guilt in this condition as inappropriate.

Table 4
Means and standard deviations for each condition (Study 4). Ratings were made on a 1–7 scale. Judgments of moral character served as our primary measure of interest.

	Guilt	*No Guilt*	*Vicarious Guilt – Near*	*Vicarious Guilt – Far*
Moral character	5.66 (0.95)	4.54[a] (1.21)	5.68 (1.14)	5.20[c] (1.18)
Social likability	5.26 (1.21)	4.24[a] (1.29)	5.12 (1.33)	4.46[c] (1.45)
Likelihood of future guilt	5.86 (1.12)	4.49[a] (1.26)	6.14 (1.04)	5.75 (1.34)
Likelihood of future gratitude	5.57 (1.14)	4.49[a] (1.34)	5.95[b] (1.06)	5.47[c] (1.16)
Blame	2.85 (1.59)	3.48[a] (1.74)	1.60[b] (1.34)	2.59 (1.84)
Likelihood of future moral offense	2.74 (1.38)	3.92[a] (1.47)	2.60 (1.63)	3.03 (1.68)
Neuroticism	3.90 (1.03)	2.93[a] (0.99)	4.95[b] (1.20)	4.85[c] (1.29)
Agent should have felt more emotion	4.17 (1.17)	4.51 (1.09)	2.35[b] (1.55)	3.04[c] (1.88)

[a] *Guilt* and *No Guilt* ratings significantly differed from each other, $p < .05$.
[b] *Guilt* and *Vicarious Guilt – Near* ratings significantly differed from each other, $p < .05$.
[c] *Guilt* and *Vicarious Guilt – Far* ratings significantly differed from each other, $p < .05$.

There was a similar pattern in terms of judgments of social likability, with an overall significant difference between conditions, $F(3, 431) = 15.24$, $p < .001$, $\eta_p^2 = 0.10$. Planned contrasts revealed a significant difference in judgments of likability between the Janet *guilt* condition and the Janet *no guilt condition*, $t(431) = 5.70$, $p < .001$, $d = 0.81$; a significant difference between the Janet *guilt* condition and the *vicarious guilt – far* condition, $t(431) = 4.44$, $p < .001$, $d = 0.60$; and no significant difference between the Janet *guilt* condition and the *vicarious guilt – near* condition, $t(431) = 0.79$, $p = .43$, $d = 0.11$.

There was an overall significant difference between conditions on predictions of the agent’s likelihood of experiencing guilt in future situations, $F(3, 431) = 40.85$, $p < .001$, $\eta_p^2 = 0.22$. Planned contrasts revealed that, compared to the Janet *guilt* condition, participants expected Janet in the *no guilt* condition to be significantly less likely to experience guilt in future situations, $t(431) = 8.50$, $p < .001$, $d = 1.15$; for Sarah in the *vicarious guilt – near* condition, $t(431) = 1.73$, $p = .09$, $d = 0.26$, and in the *vicarious guilt – far* condition, $t(431) = 0.69$, $p = .52$, $d = 0.09$, to be equally likely to experience guilt in future situations .

There was an overall significant difference between conditions on predictions of the agent’s likelihood to experience gratitude in future situations, $F(3, 431) = 30.26$, $p < .001$, $\eta_p^2 = 0.17$. Planned contrasts revealed that, compared to Janet in the *guilt* condition, participants expected Janet in the *no guilt* condition to be significantly less likely to express gratitude in other situations, $t(431) = 6.79$, $p < .001$, $d = 0.87$; expected Sarah in the *vicarious guilt – near* condition to be more likely to express gratitude in other situations, $t(431) = 2.38$, $p = .02$, $d = 0.35$; and expected Sarah in the *vicarious guilt – far* condition to be equally likely to experience gratitude in future situations, $t(431) = 0.63$, $p = .53$, $d = 0.09$.

We next examined how much blame participants assigned to the agent for what happened. We, again, found a significant overall difference between conditions, $F(3, 430) = 25.04$, $p < .001$, $\eta_p^2 = 0.15$. Compared to the *guilt* condition, participants assigned significantly more blame to the agent in the *no guilt* condition, $t(430) = 2.83$, $p = .005$, $d = 0.38$. For vicarious targets, participants assigned significantly less blame in the *vicarious guilt – near* condition, $t(430) = 5.68$, $p < .001$, $d = 0.85$, and similar amounts of blame in the *vicarious guilt—far* condition ($M =$

2.64, SD = 1.84), $t(430) = 1.17$, $p = .24$, $d = 0.15$. Consistent with our previous results, participants assigned less blame to the *guilty-feeling* Janet than to the *non-guilty-feeling* Janet for the same accident. In addition, participants blamed Sarah significantly less in the *vicarious guilt – near* condition, suggesting, perhaps, that they did not hold her morally accountable for the accident. Unexpectedly, however, there was no significant difference in the blame assigned to Janet in the *guilt* condition and Sarah in the *vicarious guilt – far* condition. Our only possible interpretation is that in this condition participants may have interpreted the question "how blameworthy is Sarah for what happened?" to refer to blame over her feelings of guilt rather than blame for the accident itself (because she obviously played no role in the events of the accident).

We next examined whether an agent's feelings (or absence of feelings) of guilt influenced judgments that the agent would commit a minor moral offense in the future. Consistent with our predictions, we found a significant overall difference between conditions, $F(3, 426) = 15.88$, $p < .001$, $\eta_p^2 = 0.10$. Contrasts revealed significantly higher judgments of likelihood in the *no guilt* condition than in the *guilt* condition, $t(426) = 5.61$, $p < .001$, $d = 0.83$. We found no significant difference in judgments of likelihood between the *guilt* condition and both the *vicarious guilt – near* condition, $t(426) = 0.69$, $p = .49$, $d = 0.09$, and the *vicarious guilt – far* condition, $t(426) = 1.36$, $p = .17$, $d = 0.19$. This suggests that it is the absence of guilt that seems to be driving predictions of future moral offenses. The mere presence of a guilt response – whether situationally true positive or not – was enough to lead to a more optimistic moral outlook when compared to an agent who expressed no guilt at all.

We also found a significant overall difference between conditions on participants judgments of Janet/Sara's dispositional neuroticism, $F(3, 431) = 75.05$, $p < .001$, $\eta_p^2 = 0.34$. Compared to the *guilt* condition, participants judged Janet as less neurotic in the *no guilt* condition, t (431) = 6.20, $p < .001$, $d = 0.96$; judged Sarah as more neurotic in the *vicarious guilt – near* condition, $t(431) = 6.81$, $p < .001$, $d = 0.94$; and judged Sarah as more neurotic in the *vicarious guilt – far* condition, $t(431) = 6.17$, $p < .001$, $d = 0.81$.

Finally, we found a significant difference of condition on participants' judgments regarding whether the agent felt the "right" amount of guilt in response to the accident, $F(3, 431) = 51.40$, $p < .001$, $\eta_p^2 = 0.26$. Compared to the *guilt* condition, observers judged that the Janet in the *no guilt* condition should have felt slightly (but non-significantly) more guilt than she did, t(431) = 1.70, $p = .09$, $d = 0.30$; in the *vicarious guilt – near* condition Sarah should have felt less guilt than she did, $t(431) = 9.21$, $p < .001$, $d = 1.33$; and that in the *vicarious guilt – far* condition Sarah should have felt less guilt than she did, $t(431) = 5.74$, $p < .001$, $d = 0.73$.

Together, our results provide evidence that observers are not simply evaluating agents based on the presence or absence of a guilt response. Instead, observers are attuned to the *appropriateness* of an individual's guilt to the situation. In these studies, observers were sensitive to whether the agent could have reasonably acted in a way that would have prevented the harm from occurring. In the absence of a reasonable counterfactual, guilt was not seen as a strong predictor of the agent's moral character.

6. Study 5

In Study 5, we sought to investigate not just the judgments that people make for agents who express "false positive" guilt (or not) over accidental harms, but to explore whether this information influences behavior toward those agents—particularly in their willingness to trust agents in a social, interactive game (the "trust" game; Berg, Dickhaut, & McCabe, 1995). We predicted that individuals would be more likely to trust an agent who displayed false positive guilt compared to an agent who did not.

6.1. Methods

6.1.1. Participants

We recruited 201 U.S. participants through MTurk. We based our sample size on those used in past research using a similar methodology (Everett, Pizarro, & Crockett, 2016). Per our preregistration, we excluded participants who failed any of our three comprehension questions regarding the trust game ($N = 52$), leaving a final sample of 149 (24% female, $M_{age} = 35.28$).

6.1.2. Design

Participants first answered open-ended questions asking how they would act in three hypothetical situations. The first situation was an adaptation of the coffee spill scenario from Studies 1–2, while the other two situations were filler tasks that were not relevant to our hypotheses.[4] The first situation read "Imagine you are in a crowded coffee shop to purchase a drink. After receiving your order, you begin making your way towards the exit. As you are walking, you fail to notice a wrapper on the floor and accidentally slip and spill your drink on someone else. How would you feel if this happened? Would you feel guilty?" Participants were then introduced to the trust game (TG). In the typical TG, there are two players: an "investor" and a "trustee." The investor is endowed with some money and told that any money they transfer (from zero to the full amount) to the trustee will be doubled, at which point the trustee can then decide to transfer a proportion of their total amount (from zero to the full amount they received) back to the investor (this amount is the measure of "trust"). After participants were given this description, they were asked three comprehension questions regarding the TG to ensure that they understood the game.

After successfully completing the comprehension questions, we then told participants that they had been assigned the role of the investor in the game, that they had been given $0.50 as their initial endowment, and that they would be playing in a trust game with one of two potential players; namely, other MTurkers who had already answered the hypothetical questions, and who had consented to sharing their answers with other participants (we reiterated that their own answers would not be shown to the other players). Participants were told that after they reported how they would behave in the trust game we would randomly select one of the other players to be the participant's partner and would carry out the decisions for real, and that the participant's final bonus payment would be based on the outcomes of these decisions.

Participants were then presented (in counterbalanced order) with the responses to the *coffee spill* scenario that had been ostensibly provided by the two other players who served as potential partners. Player 1 (*guilty*) responded by saying "Oh god, I think I would feel pretty bad about it. Even if it was an accident and it was technically not my fault, I'd feel pretty guilty." Player 2 (*non-guilty*) said "I might feel bad, but if it was an accident, why would I feel guilty? It's not like I meant to do it or anything." As an explicit measure of partner choice, we asked participants who they would most prefer as a partner in the TG, Player 1 or Player 2. As indicators of trust, we asked participants how much of their $0.50 they would want to transfer if they were playing the game with Player 1 and how much they would want to transfer if they were playing with Payer 2 (from $0.00 to $0.50), and what percentage of money they believed they would receive back if that particular player was their partner (from 0% to 100%).

[4] One filler task asked "Imagine you are walking around your town and on the sidewalk is an unmarked envelope with $100 in it. What would you do with the money?" The other filler task asked "Imagine your first cousin came to you and asked you to help cover their mortgage payment for a month. What would you do?" These filler tasks were included to increase the overall believability of the paradigm.

6.2. Results and discussion

Consistent with our hypotheses, as well as with the results from our previous studies, participants were more likely to prefer playing with the partner who reported false positive emotions (i.e., who reported that they would feel guilty in the hypothetical accident scenario; 82%) than with the partner who reported that they would not feel guilt (18%), $p <$ .001.

Because the data were non-normally distributed, we used a series of Wilcoxon signed-ranks test to compare the amount of money transferred and the percentage participants predicted they would receive in return. Supporting our hypotheses, participants transferred more money to the *guilty* partner than the *non-guilty* partner ($Z = 6.83$, $p > .001$, $r = 0.56$), and reported expecting to receive more money back from the *guilty* partner than the *non-guilty* partner ($Z = 7.61$, $p < .001$, $r = 0.62$). Together, these results provide strong evidence that people are much more trusting of others when those others experience guilt, even when the guilt is normatively unjustified.

7. Study 6

In Studies 1–5, we found that participants judge agents who feel false positive moral emotions as having a better moral character. However, it is not clear from these results whether these judgments are accurate. Is there any evidence that participants who report false positive guilt are *actually* better people? To return to Bernard Williams' example (1981), are we right to doubt the moral character of the lorry driver who is too quick to abandon his guilt over having accidentally killed someone? In Study 6, we attempted to address this question by examining whether the tendency to experience false positive moral emotions is associated with moral character using measures of character that have been developed and validated by others. Specifically, we assessed participants' empathy, aggression, callous affect, and willingness to deceive others (Paulhus & Williams, 2002) using scales of psychopathic personality, Machiavellianism, narcissism, and perceived life meaninglessness (design adapted from Bartels & Pizarro, 2011). Participants completed these individual difference measures and were asked to respond to a variety of hypothetical scenarios constructed such that a moral emotion (i.e., guilt, gratitude) was either normatively appropriate (e.g., feeling guilt when being morally responsible) or false positively appropriate (e.g., feeling guilt even when not morally responsible). We predicted that participants higher in psychopathy, narcissism, and Machiavellianism would report feeling less guilt and gratitude for both "false positive" situations and "true positive" situations (in which experiencing the emotions would be normatively appropriate). If so, we believed that this would constitute the first evidence that this tendency to over-experience moral emotions might be a reliable predictor of underlying moral character.

7.1. Methods

7.1.1. Participants

We recruited 205 U.S. participants (46% female, $M_{age} = 29.41$) through Prolific.co, an online data collection service (Palan & Schitter, 2018), and paid each $2.00 for participation. The sample size was based on previous research using a similar design (Bartels & Pizarro, 2011).

7.1.2. Design

Participants responded to four hypothetical scenarios and a battery of individual difference measures (described above). The presentation of the hypothetical scenarios and individual difference measures was counterbalanced between participants.

The hypothetical scenarios were presented in random order based on a 2 (emotion: guilt, gratitude) X 2 (appropriateness: false positive, true positive) within-subjects design. For all scenarios, we asked participants if they would feel the target emotion, either guilt or gratitude (from *1 = I would not feel guilty/grateful at all* to *7 = I would feel extremely guilty/grateful*). For each emotion (guilt and gratitude), participants read both a scenario with a false positive expression (e.g., accidentally slipping on a wrapper and spilling your coffee on someone in a coffee shop) and a different scenario with a true positive expression (e.g., unintentionally locking a visiting cousin out of the house after they left a note that they were outside).

The individual differences battery included an adapted version of a 30-item psychopathy scale with three subfactors: interpersonal manipulation, callous affect, and erratic lifestyle (SRP-III; Paulhus, Neumann, & Hare, 2009), the 18-item No Meaning scale (Kunzendorf, Moran, & Gray, 1995), the 20-item Machiavellianism scale (Mach-IV; Christie & Geis, 1970), and the Single Item Narcissism Scale (SINS; Konrath, Meier, & Bushman, 2014). We also included a 10-item social desirability scale (MC-1; Strahan & Gerbasi, 1972), a standard measure of a participant's tendency to respond in a manner that would be perceived as favorably by others. This was included in order to control for the possibility that responses to the emotional scenarios were a reflection of this tendency. Participants responded to a randomized ordering of all 79 items (from *1 = strongly disagree* to *7 = strongly agree*), including "I like to see fist-fights" (psychopathy), "When you really think about it, life is not worth the effort of getting up in the morning" (No Meaning), and "The best way to handle people is to tell them what they want to hear" (Machiavellianism). Finally, participants reported their age and gender.

7.2. Results and discussion

7.2.1. Guilt

Participants who reported feeling guilty in the false positive scenario also tended to report feeling guilty in the true positive scenario, $r(204) = 0.23$, $p = .001$, suggesting that false positive expressions of the moral emotion of guilt predict the tendency to express guilt in normatively appropriate situations and vice versa. As predicted, participants who scored higher on psychopathy ($\alpha = 0.86$; $r[205] = -0.19$, $p = .006$), Machiavellianism ($\alpha = 0.69$; $r[205] = -0.16$, $p = .02$), and narcissism ($r[204] = 0.20$, $p = .005$) reported that they would feel less guilty in the true positive scenarios compared to people who scored lower on those measures (see Table 5). However, life meaninglessness ($\alpha = 0.91$; $r[205] = -0.09$, $p = .22$) and social desirability ($r[205] = 0.05$, $p = .45$) were not significantly correlated with participants' reported guilt in the true positive scenarios. There was a similar pattern of results for reported guilt in false positive scenarios, although the effects were slightly weaker on average. The results support our primary hypothesis that moral character, as measured by individual differences in "dark triad" personality traits, is associated with the degree to which a person

Table 5
Correlations between individual difference measures (including the psychopathy subscales) and self-reported guilt and gratitude for true positive and false positive scenarios (Study 6). We were most interested in the correlations with Psychopathy, particularly the callous affect subscale.

	False positive guilt	True positive guilt	False positive gratitude	True positive gratitude
Psychopathy	$-0.13^{\dagger}$	-0.19^{**}	-0.16^{*}	-0.24^{***}
Callous affect	-0.21^{**}	-0.22^{***}	-0.21**	-0.25^{***}
Interpersonal manipulation	-0.09	$-0.13^{\dagger}$	$-0.13^{\dagger}$	-0.18^{*}
Erratic lifestyle	0.03	-0.08	-0.02	$-0.12^{\dagger}$
Machiavellianism	$-0.13^{\dagger}$	-0.16^{*}	$-0.13^{\dagger}$	-0.11
Narcissism	$-0.13^{\dagger}$	-0.20^{**}	-0.03	-0.15^{*}
No meaning	0.09	-0.09	0.01	-0.22^{**}
Social desirability	-0.01	0.05	0.04	0.06

† $p < 0.1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

experiences guilt in both false positive scenarios and true positive scenarios.

Examining the three factors of the psychopathy scale individually, we found that participants who scored higher in callous affect (α = 0.76) reported that they would feel significantly less guilt in true positive scenarios, $p = .001$, and in false positive scenarios, $p = .003$. However, there were no significant correlations between the interpersonal manipulation factor (α = 0.75) and either true positive scenario guilt, $p = .07$, or false positive guilt, $p = .22$. There was a similar lack of significant correlations between the erratic lifestyle factor (α = 0.71) and reported guilt on either true positive scenarios, $p = .27$, or false positive guilt scenarios, $p = .65$. The results from the psychopathy subscales suggest that the effects on true positive and false positive guilt are primarily driven by a tendency to experience callous affect.

Together, these results suggest that the tendency to report feeling guilt over a harmful outcome is linked to a person's degree of emotional callousness, but not necessarily their tendency to interpersonally manipulate or to have an erratic lifestyle. Overall, these results suggest that making inferences about moral character based on expressions of false positive guilt may be an accurate strategy.

7.2.2. Gratitude

Participants who reported feeling gratitude in the false positive scenario also tended to report feeling gratitude in the true positive scenario, $r(205) = 0.17$, $p = .02$, suggesting that false positive expressions of the moral emotion of gratitude do predict the tendency to express gratitude in situations that should elicit gratitude. As can be seen in Table 5, participants who scored higher on psychopathy ($p < .001$), narcissism ($p = .04$), and life meaningless ($p = .002$) predicted they would feel less gratitude in the true positive scenarios, while Machiavellianism ($p = .11$) and social desirability ($p = .36$) did not significantly correlate with predicted gratitude in the true positive scenarios. For false positive scenarios, the only significant correlation to emerge was with gratitude and psychopathy ($p = .02$). These results provide partial support for our primary hypothesis – subclinical levels of psychopathy are associated with the degree to which a person experiences gratitude in both false positive scenarios and true positive scenarios. The relative differences between predicted guilt and predicted gratitude and their associations with Machiavellianism and narcissism could be explained by the differences between guilt and gratitude, such that guilt reflects taking partial responsibility for a harmful action, responsibility that those high in Machiavellianism and narcissism may tend to avoid.

We also found that participants who scored higher in the callous affect subscale of the psychopathy measure predicted they would feel significantly less gratitude in both true positive scenarios, $p < .001$, and false positive scenarios, $p = .003$. There was also a significant correlation between the interpersonal manipulation factor and true positive scenario gratitude, $p = .01$, and a nonsignificant correlation with false positive gratitude, $p = .06$. However, there were no significant correlations between the erratic lifestyle factor and either true positive scenario gratitude, $p = .08$, or false positive gratitude, $p = .77$. Together, these results suggest that the tendency to experience gratitude is negatively linked to a person's degree of emotional callousness and their tendency to interpersonally manipulate, but not their tendency to have an erratic lifestyle. Like guilt, observers may be well-calibrated in making more favorable judgments of people who feel false positive gratitude.

8. General discussion

Collectively, our results support the hypothesis that false positive moral emotions are associated with both judgments of moral character (Studies 1–5) and traits associated with moral character (Study 6). We consistently found that observers use an agent's false positive experience of moral emotions (e.g., guilt, gratitude) to infer their underlying moral character, their social likability, and to predict both their future emotional responses and their future moral behavior. Specifically, we found that observers judge an agent who experienced "false positive" guilt (in response to an accidental harm) as a more moral person, more likeable, less likely to commit future moral infractions, and more trustworthy than an agent who experienced no guilt. Our results help explain the second "puzzle" regarding guilt for accidental actions (Kamtekar & Nichols, 2019). Specifically, one reason that observers may find an accidental agent less blameworthy, and yet still be wary if the agent does not feel guilt, is that such false positive guilt provides an important indicator of that agent's underlying character.

We find a similar effect for false positive experiences of both guilt and gratitude – an agent who experienced gratitude toward someone performing their duties was rated as having better moral character than an agent who did not experience gratitude in the same situation. Additionally, this effect was not driven by the false positive experience of emotions in general or perceived differences in overall emotionality (Study 3), or by the mere experience of guilt itself (Study 4) – observers appear to specifically infer moral character based on the false positive presence of *moral* emotions in response to actions under which the agent had reasonable control over. False positive emotions outside of the moral domain may serve as predictors to an agent's underlying disposition, but about nonmoral dispositions. For example, the presence or absence of fear when faced with harmless snakes is likely treated as a predictor of the agent's emotionality and fearlessness. In the moral domain, lay conceptions of character seem to encompass a suite of particular emotional predispositions, including the experience of false positive guilt and gratitude, the valuing of individual lives (Everett et al., 2016), and the experience of "warm glow" emotions after prosocial behavior (Barasch, Levine, Berman, & Small, 2014).

We also demonstrated that these inferences of character have implications for how individuals behave toward an agent. Specifically, agents who report anticipating guilt for an accident were trusted more and were more likely to be preferred as an interaction partner than agents who do not (Study 5). These findings extend a growing body of research on the behavioral predictors of trustworthiness, adding to a list that includes a willingness to make intuitive, deontological moral judgments (Everett et al., 2016), cooperating without carefully calculating costs and benefits (Jordan, Hoffman, Nowak, & Rand, 2016), and a willingness to engage in third-party punishment (Jordan, Hoffman, Bloom, & Rand, 2016).

Finally, we found that inferences of moral character from an agent's false positive moral emotions may actually be warranted. In Study 6, we showed that participants who scored higher on measures of psychopathy, Machiavellianism, and Narcissism reported that they would feel less guilt in response to accidental harms and less gratitude toward someone who helped them, compared to participants who scored lower on those measures. This association between these "dark triad" personality traits (Paulhus & Williams, 2002) and reported moral emotions held for both true positive cases (i.e., situations where guilt and gratitude would be normatively appropriate) and false positive cases (i.e., situations where guilt and gratitude would not necessarily be normatively appropriate).

8.1. Moral emotions as predictors

These findings make sense given the body of research that has focused on the social function of moral emotions (Algoe & Haidt, 2009; Haidt, 2003; Hutcherson & Gross, 2011; Tangney, Stuewig, & Mashek, 2007). For example, feelings of guilt can motivate attempts to repair a damaged relationship (Schmader & Lickel, 2006; Tangney, Miller, Flicker, & Barlow, 1996; Wicker, Payne, & Morgan, 1983). Moreover, agents who anticipate feeling aversive emotions like guilt and regret for a decision tend to avoid making that decision (e.g., Massi Lindsey, 2005; Steenhaut & Van Kenhove, 2006). Guilt seems to serve a self-regulatory function, modulating behavior to discourage cheating and other norm violations and making someone a better cooperative partner (Frank, 1988; Prinz, 2004; Trivers, 1971).

It seems reasonable to think that there would be some benefit to communicating these moral emotions as a signal of character, and to being able to glean information about the character of others from observations of their emotional responses. If a propensity to feel guilt makes it more likely that a person is cooperative and trustworthy, observers would need to discriminate between people who are and are not prone to guilt. Guilt could therefore serve as an effective regulator of moral behavior in others in its role as a reliable signal of good character. This account is consistent with theoretical accounts of emotional expressions more generally, either in the face, voice, or body, as a route by which observers make inferences about a person's underlying dispositions (Frank, 1988). Our results suggest that false positive emotional responses specifically may provide an additional, and apparently informative, source of evidence for one's propensity toward moral emotions and moral behavior.

Our results can also provide insight into understanding collective guilt (i.e., guilt in response to harm done by members of one's ingroup; Wohl, Branscombe, & Klar, 2006). Observers could treat an individual's guilt for a collective action as a false positive expression of guilt (given that the individual is not causally responsible for the actions for their group members and blame is thus normatively inappropriate), and therefore judge the individual as a more moral person. There may be merit to this inference, as collective guilt has been linked with support for policies that address group inequities (Brown, González, Zagefka, Manzi, & Čehajić, 2008).

8.2. Limitations and future directions

Our studies provide an initial examination of the role of false positive moral emotions in judgments of moral character. Of course, additional work is necessary to replicate and extend our findings, as well as to address potential limitations. One such limitation is that our studies utilized several single-item measures, which may have reliability issues (e.g., Wanous & Reichers, 1996), so future work should aim to replicate a more robust set of measures. Furthermore, while we used a variety of different vignettes and methods (e.g., the trust game in Study 5), our account would benefit from additional work that used non-vignette methods to examine how observers react in situ to someone expressing guilt for a real accident. In addition, several of our studies used a within-subjects design – while such designs often increase statistical power, they may inadvertently increase the likelihood of suspicion and demand responses (for a review of the differences in these designs, see Charness, Gneezy, & Kuhn, 2012).

While we have provided an initial examination of guilt and gratitude, we believe there are open questions regarding both emotions and their connection to perceived moral character. For example, interpreting guilt as "false positive" could depend on whether an accident is *unforeseeable* (i.e., the agent could not have knowledge of the potential harmful consequences), or *foreseeable but unforeseen due to negligence* (i.e., the agent could have foreseen the harm if they had been more vigilant and attentive). Our studies also leave open the question of whether the experience of false positive guilt is perceived as a positive indicator of character, or whether the lack of experiencing guilt is perceived as a negative indicator of character.

Our studies were inspired by cases like Williams' lorry driver (1981), and the experiences detailed on accidentalimpacts.org, where the primary emotion of interest is guilt. Across our studies, we provided evidence that people use false positive responses of both guilt and gratitude to infer moral character. But guilt and gratitude are a small slice of the full range of moral emotions. It will be important to see whether people also perceive false positive emotional responses of other moral emotions as similar predictor of moral character, such as shame (e.g., Niedenthal et al., 1994), embarrassment (e.g., Tangney et al., 1996), anger (e.g., Russell & Giner-Sorolla, 2011), and disgust (e.g., Giner-Sorolla & Chapman, 2017; but for a critique of the usefulness of "moral disgust" as a concept, see Landy & Piazza, 2019). Because our studies revealed that false positive experiences of guilt and gratitude gave rise to judgments of moral character, but experiences of fear (a non-moral emotion in this context) did not, we would hypothesize that these effects might generalize to moral emotions. However, future research would be needed to directly test this claim with other false positive moral emotions. It is possible, for example, that moral anger has a more complicated connection to judgments of moral character than guilt, as there might be social pressure to minimize false positive anger and moral condemnation because of the potential costs of misapplied anger (e.g., resentment and retaliation; Aquino, Tripp, & Bies, 2001; McCullough, Kurzban, & Tabak, 2013).

One clear limitation of these studies is that our samples were exclusively drawn from U.S. populations using online recruitment methods, limiting our ability to generalize our findings to other populations (especially when it comes to the correlational findings from Study 6). For instance, it is known that online convenience samples may differ in important ways from random samples of the nation's general population (e.g., Arditte, Çek, Shaw, & Timpano, 2016). In addition, researchers have documented differences in norms regarding the experience and expression of emotions across cultures (Mesquita, 2001; Tracy & Robins, 2007; Tsai, Knutson, & Fung, 2006). Specifically, for the purposes of our hypotheses, there has been research showing cultural variability in the sorts of circumstances that reliably trigger both guilt (e.g., Bear, Uribe-Zarain, Manning, & Shiomi, 2009; Onwezen, Bartels, & Antonides, 2014; Stipek, 1998) and gratitude (e.g., Morgan, Gulliford, & Kristjánsson, 2014; Naito, Wangwan, & Tani, 2005). What counts as a "false positive" moral emotion is likely a contextualized judgment that varies reliably based on the particular culture being studied. Ideally, future research would address this by using a variety of tools to collect U.S. samples, and by extending the collection of data to non-U.S. populations.

Finally, the tendency to infer moral character from an agent's false positive moral emotions likely requires an understanding of the situational (i.e., when such emotions typically occur) and cultural norms (i.e., the particular display rules and cultural valuation of those emotions) for those emotions. Therefore, one possibility is that the tendency to infer moral character from an agent's false positive moral emotions develops later in life, after children have received enough input regarding those norms. Research has demonstrated that age plays a role in a variety of moral judgments (e.g., Heiphetz, Strohminger, Gelman, & Young, 2018; McAuliffe, Raihani, & Dunham, 2017), and that some of these differences are not merely due to age-related differences in cognitive ability (e.g., Starmans & Bloom, 2016) but instead to differences in exposure and learning. Future research investigating the role of exposure to moral and emotional norms on judgments of moral character could help shed light on this question.

9. Conclusion

We have provided evidence that observers use the experience of false positive moral emotions as predictors of an agent's underlying moral character, and as a way to predict an agent's future moral behavior. We have also provided initial evidence that individuals who report that they would experience false positive moral emotions may actually be more likely to possess good moral character. This research may help to understand cases in which observers blame agents very little for their accidents, yet prefer those agents to feel guilty. More broadly, our findings highlight the importance of emotions and emotional reactions in people's conceptions of what it means to be a "good person."

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104770.

Open practices

All preregistration documentation, materials, data, and analysis scripts for these studies are available on the Open Science Framework at https://osf.io/btwsq/

Declaration of Competing Interest

None.

Acknowledgements

This work was presented as a talk at multiple academic conferences, including at SPSP 2020 and at Yale University. We would like to thank our audience members for their helpful comments and feedback during these presentations.

References

Algoe, S. B., & Haidt, J. (2009). Witnessing excellence in action: The “other-praising” emotions of elevation, gratitude, and admiration. *The Journal of Positive Psychology, 4* (2), 105–127. https://doi.org/10.1080/17439760802650519.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63* (3), 368–378. https://doi.org/10.1037/0022-3514.63.3.368.

Ames, D. R., & Johar, G. V. (2009). I’ll know what you’re like when I see how you feel: How and when affective displays influence behavior-based impressions. *Psychological Science, 20*(5), 586–593. https://doi.org/10.1111/j.1467-9280.2009.02330.x.

Aquino, K., Tripp, T. M., & Bies, R. J. (2001). How employees respond to personal offense: The effects of blame attribution, victim status, and offender status on revenge and reconciliation in the workplace. *Journal of Applied Psychology, 86*(1), 52–59. https://doi.org/10.1037/0021-9010.86.1.52.

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment, 28*(6), 684–691. https://doi.org/10.1037/pas0000217.

Armsby, R. E. (1971). A reexamination of the development of moral judgments in children. *Child Development, 42*, 1241–1248. https://doi.org/10.2307/1127807.

Badar, M. E., & Marchuk, I. (2013). A comparative study of the principles governing criminal responsibility in the major legal systems of the world (England, United States, Germany, France, Denmark, Russia, China, and Islamic legal tradition). *Criminal Law Forum, 24*, 1–48. https://doi.org/10.1007/s10609-012-9187-z.

Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology, 107*(3), 393–413. https://doi.org/10.1037/a0037207.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*(1), 154–161. https://doi.org/10.1016/j.cognition.2011.05.010.

Bear, G. G., Uribe-Zarain, X., Manning, M. A., & Shiomi, K. (2009). Shame, guilt, blaming, and anger: Differences between children in Japan and the US. *Motivation and Emotion, 33*(3), 229–238. https://doi.org/10.1007/s11031-009-9130-8.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122–142. https://doi.org/10.1006/game.1995.1027.

Brandt, M. J., & Reyna, C. (2011). The chain of being: A hierarchy of morality. *Perspectives on Psychological Science, 6*(5), 428–446. https://doi.org/10.1177/1745691611414587.

Brown, R., González, R., Zagefka, H., Manzi, J., & Čehajić, S. (2008). Nuestra culpa: Collective guilt and shame as predictors of reparation for historical wrongdoing. *Journal of Personality and Social Psychology, 94*(1), 75–90. https://doi.org/10.1037/0022-3514.94.1.75.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 16. https://doi.org/10.5334/joc.72.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009.

Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism*. New York: Academic Press.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308–315. https://doi.org/10.1177/1948550612457688.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006.

Darley, J. M., Klosson, E. C., & Zanna, M. P. (1978). Intentions and their contexts in the moral judgments of children and adults. *Child Development, 49*(1), 66–74. https://doi.org/10.2307/1128594.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology, 41*, 525–556. https://doi.org/10.1146/annurev.ps.41.020190.002521.

Davis, C. G., Lehman, D. R., Wortman, C. B., Silver, R. C., & Thompson, S. C. (1995). The undoing of traumatic life events. *Personality and Social Psychology Bulletin, 21*(2), 109–124. https://doi.org/10.1177/0146167295212002.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772–787. https://doi.org/10.1037/xge0000165.

Fedotova, N. O., Fincher, K. M., Goodwin, G. P., & Rozin, P. (2011). How much do thoughts count? Preference for emotion versus principle in judgments of antisocial and prosocial behavior. *Emotion Review, 3*, 316–317. https://doi.org/10.1177/1754073911402387.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Frank, R. (1988). *Passions within reason: The strategic role of the emotions*. New York, NY: Norton.

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond purity: Moral disgust toward bad character. *Psychological Science, 28*(1), 80–91. https://doi.org/10.1177/0956797616673193.

Gold, G. J., & Weiner, B. (2000). Remorse, confession, group identity, and expectancies about repeating a transgression. *Basic and Applied Social Psychology, 22*(4), 291–300. https://doi.org/10.1207/15324830051035992.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148–168. https://doi.org/10.1037/a0034726.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford: Oxford University Press.

Haney, C., Sontag, L., & Constanzo, S. (1994). Deciding to take a life: Capital juries, sentencing instructions, and the jurisprudence of death. *Journal of Social Issues, 50* (2), 149–176. https://doi.org/10.1111/j.1540-4560.1994.tb02414.x.

Hartley, A. G., Furr, R. M., Helzer, E. G., Jayawickreme, E., Velasquez, K. R., & Fleeson, W. (2016). Morality’s centrality to liking, respecting, and understanding others. *Social Psychological and Personality Science, 7*(7), 648–657. https://doi.org/10.1177/1948550616655359.

Heiphetz, L., Strohminger, N., Gelman, S. A., & Young, L. L. (2018). Who am I? The role of moral beliefs in children’s and adults’ understanding of identity. *Journal of Experimental Social Psychology, 78*, 210–219. https://doi.org/10.1016/j.jesp.2018.03.007.

Helzer, E. G., & Critcher, C. R. (2018). What do we evaluate when we evaluate moral character? In K. Gray, & J. Graham (Eds.), *Atlas of moral psychology* (pp. 99–107). New York: Guilford Press.

Higgins, E. T. (1998). The aboutness principle: A pervasive influence on human inference. *Social Cognition, 16*(1), 173–198. https://doi.org/10.1521/soco.1998.16.1.173.

Hoffman, M. L. (1982). Development of prosocial motivation: Empathy and guilt. In N. Eisenberg (Ed.), *The development of prosocial behavior* (pp. 281–313). San Diego, CA: Academic Press.

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology, 100*(4), 719–737. https://doi.org/10.1037/a0022408.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473–476. https://doi.org/10.1038/nature16981.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences, 113*(31), 8658–8663. https://doi.org/10.1073/pnas.1601280113.

Kamtekar, R., & Nichols, S. (2019). Agent-regret and accidental agency. In *, 43. Midwest studies in philosophy* (pp. 181–202). https://doi.org/10.1111/misp.12112.

Konrath, S., Meier, B. P., & Bushman, B. J. (2014). Development and validation of the Single Item Narcissism Scale (SINS). *PLoS ONE, 9*(8). https://doi.org/10.1371/journal.pone.0103469.

Kunzendorf, R. G., Moran, C., & Gray, R. (1995). Personality traits and reality-testing abilities, controlling for vividness of imagery. *Imagination, Cognition and Personality, 15*(2), 113–131. https://doi.org/10.2190/B76E-MJ9E-07AV-KAKK.

Landy, J. F., & Piazza, J. (2019). Reevaluating moral disgust: Sensitivity to many affective states predicts extremity in many evaluative judgments. *Social Psychological and Personality Science, 10*(2), 211–219. https://doi.org/10.1177/1948550617736110.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*(2), 101–121. https://doi.org/10.1006/jesp.1996.1314.

Mandel, D. R., & Dhami, M. K. (2005). “What I did” versus “what I might have done”: Effect of factual versus counterfactual thinking on blame, guilt, and shame in prisoners. *Journal of Experimental Social Psychology, 41*(6), 627–635. https://doi.org/10.1016/j.jesp.2004.08.009.

Massi Lindsey, L. L. (2005). Anticipated guilt as behavioral motivation: An examination of appeals to help unknown others through bone marrow donation. *Human Communication Research, 31*(4), 453–481. https://doi.org/10.1093/hcr/31.4.453.

McAuliffe, K., Raihani, N. J., & Dunham, Y. (2017). Children are sensitive to norms of giving. *Cognition, 167*, 151–159. https://doi.org/10.1016/j.cognition.2017.01.006.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences, 36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160.

McCullough, M. E., Kilpatrick, S. D., Emmons, R. A., & Larson, D. B. (2001). Is gratitude a moral affect? *Psychological Bulletin, 127*(2), 249–266. https://doi.org/10.1037/0033-2909.127.2.249.

Mesquita, B. (2001). Emotions in collectivist and individualist contexts. *Journal of Personality and Social Psychology, 80*(1), 68–74. https://doi.org/10.1037/0022-3514.80.1.68.

Morgan, B., Gulliford, L., & Kristjánsson, K. (2014). Gratitude in the UK: A new prototype analysis and a cross-cultural comparison. *The Journal of Positive Psychology, 9*(4), 281–294. https://doi.org/10.1080/17439760.2014.898321.

Naito, T., Wangwan, J., & Tani, M. (2005). Gratitude in university students in Japan and Thailand. *Journal of Cross-Cultural Psychology, 36*(2), 247–263. https://doi.org/10.1177/0022022104272904.

Niedenthal, P. M., Tangney, J. P., & Gavanski, I. (1994). "If only I weren't" versus" If only I hadn't": Distinguishing shame and guilt in counterfactual thinking. *Journal of Personality and Social Psychology, 67*(4), 585–595. https://doi.org/10.1037/0022-3514.67.4.585.

Onwezen, M. C., Bartels, J., & Antonides, G. (2014). Environmentally friendly consumer choices: Cultural differences in the self-regulatory function of anticipated pride and guilt. *Journal of Environmental Psychology, 40*, 239–248. https://doi.org/10.1016/j.jenvp.2014.07.003.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004.

Paulhus, D. L., Neumann, C. F., & Hare, R. D. (2009). *Manual for the self-report psychopathy scale*. Toronto: Multi-Health Systems.

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36*(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6.

Perkins, R. M. (1939). A rationale of mens rea. *Harvard Law Review, 52*, 905–928. https://doi.org/10.2307/1334184.

Perugini, M., & Bagozzi, R. P. (2004). The distinction between desires and intentions. *European Journal of Social Psychology, 34*(1), 69–84. https://doi.org/10.1002/ejsp.186.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise. *Psychological Science, 14*(3), 267–272. https://doi.org/10.1111/1467-9280.03433.

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer, & P. R. Shaver (Eds.), *Herzliya series on personality and social psychology. The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). American Psychological Association. https://doi.org/10.1037/13091-005.

Prinz, J. (2004). Which emotions are basic? In P. Cruise, & D. Evans (Eds.), *Emotion, evolution and rationality* (pp. 69–87). Oxford, UK: Oxford University Press.

Raz, J. (2010). Being in the world. *Ratio, 23*(4), 433–452. https://doi.org/10.1111/j.1467-9329.2010.00477.x.

Royzman, E., & Kumar, R. (2004). Is consequential luck morally inconsequential? Empirical psychology and the reassessment of moral luck. *Ratio, 17*(3), 329–344. https://doi.org/10.1111/j.0034-0006.2004.00257.x.

Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. *Emotion, 11*(2), 233. https://doi.org/10.1037/a0022598.

Schmader, T., & Lickel, B. (2006). The approach and avoidance function of guilt and shame emotions: Comparing reactions to self-caused and other-caused wrongdoing. *Motivation and Emotion, 30*(1), 42–55. https://doi.org/10.1007/s11031-006-9006-0.

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

Sher, G. (2009). *Who knew?*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195389197.001.0001.

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development, 57*(1). https://doi.org/10.2307/1130649.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). In D. Bartels, C. W. Bauman, L. J. Skitka, & D. Medin (Eds.), *Moral judgment and decision making: The psychology of learning and motivation* (Vol 50). San Diego, CA: Elsevier.

Smith, R. H., Webster, J. M., Parrott, W. G., & Eyre, H. L. (2002). The role of public exposure in moral and nonmoral shame and guilt. *Journal of Personality and Social Psychology, 83*(1), 138–159. https://doi.org/10.1037/0022-3514.83.1.138.

Sperber, D. (1996). *Explaining culture*. Oxford: Blackwell press.

Starmans, C., & Bloom, P. (2016). When the spirit is willing, but the flesh is weak: Developmental differences in judgments about inner moral conflict. *Psychological Science, 27*(11), 1498–1506. https://doi.org/10.1177/0956797616665813.

Steenhaut, S., & Van Kenhove, P. (2006). The mediating role of anticipated guilt in consumers' ethical decision-making. *Journal of Business Ethics, 69*(3), 269–288. https://doi.org/10.1007/s10551-006-9090-9.

Stipek, D. (1998). Differences between Americans and Chinese in the circumstances evoking pride, shame, and guilt. *Journal of Cross-Cultural Psychology, 29*(5), 616–629. https://doi.org/10.1177/0022022198295002.

Strahan, R., & Gerbasi, K. C (1972). Short, homogeneous versions of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology, 28*(2), 191–193. https://doi.org/10.1002/1097-4679(197204)28:2&lt;191::AID-JCLP2270280220&gt;3.0.CO;2-G.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005.

Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. B. (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology, 70*(6), 1256–1269. https://doi.org/10.1037/0022-3514.70.6.1256.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145.

Tracy, J. L., & Robins, R. W. (2007). The psychological structure of pride: A tale of two facets. *Journal of Personality and Social Psychology, 92*(3), 506–525. https://doi.org/10.1037/0022-3514.92.3.506.

Tracy, J. L., & Robins, R. W. (2008). The automaticity of emotion recognition. *Emotion, 8*(1), 81–95. https://doi.org/10.1037/1528-3542.8.1.81.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*(1), 35–57. https://doi.org/10.1086/406755.

Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology, 90*(2), 288–307. https://doi.org/10.1037/0022-3514.90.2.288.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81. https://doi.org/10.1177/1745691614556679.

Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. Oxford: Oxford University Press.

Wanous, J. P., & Reichers, A. E. (1996). Estimating the reliability of a single-item measure. *Psychological Reports, 78*(2), 631–634. https://doi.org/10.2466/pr0.1996.78.2.631.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548.

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

Weiner, B., Graham, S., & Chandler, C. (1982). Pity, anger, and guilt: An attributional analysis. *Personality and Social Psychology Bulletin, 8*(2), 226–232. https://doi.org/10.1177/0146167282082007.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10*(3), 390–399. https://doi.org/10.1177/1745691614564879.

Wicker, F. W., Payne, G. C., & Morgan, R. D. (1983). Participant descriptions of guilt and shame. *Motivation and Emotion, 7*(1), 25–39. https://doi.org/10.1007/BF00992963.

Williams, B. (1981). *Moral luck: Philosophical papers 1973–1980*. Cambridge University Press.

Wohl, M. J. A., Branscombe, N. R., & Klar, Y. (2006). Collective guilt: Emotional reactions when one's group has done wrong or been wronged. *European Review of Social Psychology, 17*, 1–37. https://doi.org/10.1080/10463280600574815.