MORAL REACTIONS AS MORAL SIGNALS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Rajen Alexander Anderson

August 2021

MORAL REACTIONS AS MORAL SIGNALS

Rajen Alexander Anderson, Ph. D.

Cornell University 2021

The work presented here encompasses two lines of research broadly concerned with understanding the inferences people make from observing the moral reactions of others. The first line of research is concerned with the role that moral praise may play in signaling normative boundaries. I examined two foundational questions about moral praise. First, what makes an action praiseworthy? One possibility is the normative exceptionality, or supererogatory nature, is what makes certain actions praiseworthy. I found that participants reported actions that exceed duties (compared to duties) deserve greater praise, are less likely to happen, and are more likely to be directed at social targets. Second, what do observers infer from praise? Praise may communicate information about both the norms of a society (e.g., the action is uncommon) and about the person giving the praise (e.g., the person thinks that the action is uncommon). I found that participants inferred that praised moral behavior is less common in society, is less required and expected of people, and that the praise-giver would want to be friends with the praised agent. These studies provide insight into the process of giving praise and how praise can signal moral norms regarding duties and expectations.

The second line of research is concerned with inferences made from others' emotional expressions. 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, I found 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 guilt that accompanies it) would thus be an inappropriate response. I found that observers rate an agent who feels guilt, compared to a control agent, as a better person, less blameworthy for the accident, and less likely to commit moral offenses.

BIOGRAPHICAL SKETCH

Rajen Anderson was born in Pennsylvania, but then moved to and grew up in Orlando, FL. Afraid of all the hurricanes, tornadoes, volcanoes, and earthquakes he was sure occurred in Florida, he eventually warmed up to his new home after learning that only two of those fears were true. At different times during his childhood and adolescence, his career aspirations included being a paleontologist, a video game designer, a pharmacist, an astrophysicist, and a biomedical researcher. However, those were all put aside after he took AP Psychology in high school and vowed that research psychology was for him. At the University of Florida, he joined the research lab of Dr. James Shepperd after taking his Social Psychology course. After graduating, Raj worked for a year in the healthcare industry before enrolling in the Psychology MA program at Wake Forest University. Completing his slow journey back to the north, he then continued his psychology adventures at Cornell University. After graduation, he will stay in the north and be a postdoctoral fellow at the Kellogg School of Management at Northwestern University.

ACKNOWLEDGMENTS

Words cannot adequately express how much gratitude and appreciation I feel to the people I want to acknowledge, but they will have to work for here and now.

First, I would like to thank my parents, David Anderson and Mamata Patnaik, for all of their support and encouragement as I have been a student perhaps much longer than they anticipated. I truly would not be the person I am without them. To my sister, Jayashree, for being a true friend, despite all her teasing of me being a “punk” and a “dork”. To my grandparents, Amiya and Kabita Patnaik (Aja and Aie) – if not for them leaving India and bringing the family to the U.S. for education, I almost certainly would not exist, let alone have the opportunities I have had.

Second, I would like to thank my committee. A huge debt is owed to David Pizarro: for being the best advisor I could have asked for; for his humor, candor, kindness, and intelligence; for buying me snacks and drinks from Ithaca Coffee Company; for his trust and support of me and my ideas; for teaching me how to think about psychology in a question-driven, holistic manner. I have grown greatly as a thinker and a person by working with and knowing him. I would also like to thank my unofficial second advisor Katherine Kinzler: for teaching me about development and how to trace cognition from infancy to adulthood; for demonstrating the importance of compassion, rigor, persistence, and determination in research and life; for her consistent encouragement of me and my successes. To Tom Gilovich: I feel like I learned something new and profound about social psychology, and therefore of life, from every class, prosem, talk, and conversation I have had the privilege of sharing with him. To Melissa Ferguson: for being a true role model, in every sense of the term, for me, other students, and the field. To Michael Goldstein: for embodying the values of the interdisciplinary pursuit of knowledge, for teaching me how to write grants, and showing the power of comparative research.

To the other Cornell faculty I have been lucky to collaborate with – David Dunning, Amy Krosch, Shaun Nichols and Rachana Kamtekar – I am thankful for the opportunity to learn from and work with them. The projects reported here are in collaboration with Shaun (Chapters 2 and 3) and Rachana (Chapter 3). When I talk about Cornell (especially to prospective graduate students), I always emphasize the diversity of experience afforded to us. I am grateful that I have been able to see so many different approaches and styles to research from so many amazing people.

To my previous advisors, James Shepperd at the University of Florida and E. J. Masicampo at Wake Forest University: thank you believing in me and teaching an eager student how to turn nascent questions into testable experiments.

A big shout out to my fellow graduate students at Cornell, both in psychology and out. There are so many bright, talented, hard-working people in the department, and it has been amazing to be a part of that group. A special thanks to Benjamin Ruisch, for being one of the best friends I have ever made and for being one of the best collaborators to work with. I look forward to all of our future projects, buddy.

Thank you to the many Cornell undergraduates I have had the opportunity to talk to, teach, and mentor: Sarene Shaked, Darby Tarlow, Anisha Duvvi, Nina Oleynikov, and many more. Much of my development as an academic at Cornell has come from working with you all.

Finally, to my partner Yasemin Kalender, who has shown me more love, care, support, and understanding than I probably deserve. I am amazed that I found someone as intelligent, warm, strong, and funny as she is. I knew I *could* succeed because she knew I *would* succeed. The final years of my PhD, and the past year of the COVID-19 pandemic, have been so much brighter having her in my life. I cannot wait for all of our upcoming adventures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

One of the most fundamental dimensions by which humans evaluate and organize their social world is based on moral qualities (Critcher et al., 2020; Fiske, 2018; Goodwin et al., 2014; Hartley et al., 2016; Pizarro & Tannenbaum, 2012; Uhlmann et al., 2015). Morality is frequently defined in psychology as the social norms, customs, and cognitions that facilitate self-control and sustain social relationships in order to promote cooperation among group members (e.g., Curry, 2016; Greene, 2015; Haidt, 2008; Haidt & Kesebir, 2010; Janoff-Bulman & Carnes, 2013; Joyce, 2006; Rai & Fiske, 2011; Sterelny & Fraser, 2016; Tomasello & Vaish, 2013). To accomplish these functions, our moral systems employ a variety of tools, including affectively-intense prescriptive rules and norms about how people should and should not behave (Nichols, 2002), emotions (e.g., guilt and gratitude) to motivate behavior (Haidt, 2003; Tangney, Stuewig, & Mashek, 2007), and social judgments of responsibility (e.g., blame and praise) to regulate behavior (Malle et al., 2014).

The present projects aimed to help address two important questions regarding moral psychology. The first question (Chapter 2) is concerned with how people come to learn and understand moral norms regarding a particular behavior. Broadly defined, moral norms are shared notions about what is right and wrong with respect to how people should treat each other (Harms & Skyrms, 2008). Norms, moral or otherwise, can have a strong impact on people's behavior and judgments (e.g., Cialdini, 2003; Crandall et al., 2018; Gelfand et al., 2017; McAuliffe et al., 2017; Paluck & Shepherd,

2012; Reno et al., 1993). Research on norms frequently distinguishes between injunctive or prescriptive norms (i.e., what people should do) and descriptive norms (i.e., what people typically do), and people learn about and respond to both types of norms (Bear & Knobe, 2017; Cialdini et al., 1990; Eriksson et al., 2015; Lapinski & Rimal, 2005). The particular content of those norms differs across people and across cultures (Awad et al., 2020; Graham et al., 2009; Jonathan Haidt & Graham, 2007; Miller et al., 1990; Rai & Fiske, 2011; Shweder et al., 1987). However, for purposes of the present work, I am agnostic regarding *which* particular moral norms that people learn, focusing instead on *how* people learn such prosocial norms. That is, what is a general mechanism by which people understand the surrounding cultural and situational norms regarding prosocial behavior?

The second question (Chapter 3) is concerned with how people evaluate the moral character of others. How do people determine whether someone is a trustworthy interaction partner? Beyond just judgments of actions, research in moral psychology has increasingly highlighted the importance of understanding a person's character (Critcher et al., 2020; Hartley et al., 2016; Helzer & Critcher, 2018; Uhlmann et al., 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 et al., 2014) and define personal identity largely in moral terms (Heiphetz et al., 2018; Strohminger & Nichols, 2014). Critical for the evaluation of character is determining whether someone is likely to cooperate and help or potentially cheat and injure. Our ability to cooperate with each other – to put aside selfishness in favor of trust and generosity – is often cited as the key to our

success as a species (Axelrod & Hamilton, 1981; Nowak, 2006). However, part of this success depends on individuals being able to identify *who* is worth cooperating with (Cosmides, 1989; DeSteno et al., 2012; Frank et al., 1993; Kinzler & Shutts, 2008; Rand & Nowak, 2013). By what methods do people judge the moral character of others?

In the present work, I provide evidence for one solution to each of these questions. Like many aspects of our psychological lives, a perhaps obvious starting point for how people attempt to solve these two dilemmas is social information and learning from and observing others (e.g., Bandura, 1977; Bronfenbrenner, 1986; Goldstein & Schwade, 2008; Jordan, Hoffman, Nowak, et al., 2016; Martin et al., 2017). Other people serve as a rich source of information, both about themselves and the greater social world. Specifically, I investigate one particular channel of social information: that of people's morally-relevant *reactions* and *responses*. I argue that one method by which humans learn about prosocial norms and the moral character of others is by observing how other people react and respond to (e.g., their reflexive judgments and emotions) morally-relevant events and triggers (e.g., someone's moral behavior). The philosopher P. F. Strawson called such responses "reactive attitudes" (1962), encompassing moral judgments (e.g., praise and blame) and emotions (e.g., guilt, gratitude, anger, and sympathy).

Why would people use such reactions in their own judgments and attitudes? Most importantly, people frequently believe that such reactions *mean* something. People believe that thoughts and other cognitions, especially their own, reflect the true state of the world (Gilbert, 1991; Griffin & Ross, 1991; Pronin et al., 2004).

Therefore, by observing how another person thinks about a moral event, people can potentially gain information about how others may think about that event and how that initial target may think about other moral events and their own potential moral behavior. For example, observers treat an agent's condemnation of a behavior as indication that the agent would be unlikely to engage in that same behavior (Hok et al., 2020; Jordan et al., 2017; Jordan, Hoffman, Bloom, et al., 2016). In addition, observers use a person's emotional expressions as indicative of what that person is like (Ames & Johar, 2009; Dijk et al., 2009, 2011; Martin et al., 2017). In the chapters that I follow, I focus on two specific instances of this more general principle.

In Chapter 2, I explore moral praise as one such moral response and the role praise can play in providing information regarding prosocial norms. Moral praise is the expression that a person has done something morally good or is a good person. In six studies, I examined two foundational questions about moral praise. First, what makes an action praiseworthy? I investigated one possibility: it is the normative exceptionality, or supererogatory nature, that makes certain actions praiseworthy. In Study 1, I found that participants reported actions that exceed duties (compared to duties) deserve greater praise, are less likely to happen, and are more likely to be directed at social targets. Having established what quality appears to make actions praiseworthy, I turn to the second question: what do observers infer from praise? Praise may communicate information about both the norms of a society (e.g., the action is uncommon) and about the person giving the praise (e.g., the person thinks that the action is uncommon). In Studies 2-5, participants read vignettes that depicted one person praising another person – compared to acknowledgment of the behavior

(Studies 2-4), gratitude for the behavior (Studies 3a and 3b), and blame of the behavior (Study 5). I found that participants inferred that praised moral behavior is less common in society, is less required and expected of people, and that the praise-giver would want to be friends with the praised agent. These studies provide insight into the process of giving praise and how praise can signal moral norms regarding duties and expectations. Chapter 2 is a reproduction of a paper, in collaboration with Shaun Nichols and David Pizarro, that is under review at the *Journal of Personality and Social Psychology*.

In Chapter 3, I investigate how moral emotions (primarily guilt, but also gratitude) in response to false positive triggers (i.e., events that theoretically should not trigger the emotion but may nonetheless do so descriptively) are treated by observers as signals to the agent's moral character. For example, people often feel guilt for accidents—negative events that they did not intend or have any control over. Why might this be the case? Across six studies, I find support for the hypothesis that observers expect "false positive" emotions from agents during a moral encounter. 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 I 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, I 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, I find that observers are more willing to trust others who experience false positive guilt compared to those who do not. In Study 6, I 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. Chapter 3 is a slightly edited version of a paper, in collaboration with Rachana Kamtekar, Shaun Nichols, and David Pizarro, that is in press at *Cognition*.

A quick note. I have elected to use the pronoun "I" throughout, when in reality both projects were very much "We" endeavors. Any praise for the works should be directed to my collaborators (David, Shaun, and Rachana), while any blame should be mine alone (Schein et al., 2020).

CHAPTER 2

PRAISE IS FOR ACTIONS THAT ARE NEITHER EXPECTED NOR REQUIRED

In the present set of studies, I examined the interplay between moral norms and moral evaluations, seeking to address two questions regarding moral praise. First, what sorts of actions generate praise? That is, beyond being "morally good", is there some unifying dimension that connects actions that are praiseworthy? I investigated one possibility: it is the normative exceptionality, or supererogatory nature, of an action that makes certain actions praiseworthy. That is, praise is characteristically judged to be appropriate when an action exceeds an agent's duties. Second, what inferences do people make from observing judgments of praise? Extending the first question, when observers see that an action was praised, do they then infer that the action was exceptional and supererogatory? One potential function of praise may be to not simply act as a direct reward learning mechanism but also to signal what is and is not normative and expected.

Moral Praise

In philosophy, praise has been characterized as indicating that an action is "laudable" (Smith, 1991) and that agents are praiseworthy when performing morally good actions for morally worthy motives (Arpaly & Schroeder, 1999). Building on the conceptual work in philosophy, by moral praise, I mean "a cognitive appraisal regarding an agent's positive moral behavior and character" (Anderson et al., 2020, p. 694). However, although actions and agents may be privately deemed praiseworthy, praise is characteristically a public expression. That is, to "praise someone" typically

means to *communicate* the appraisal that a target possesses positive moral character or has performed a moral action. When praise is thus communicated, people can draw inferences from those public expressions. Although my design focuses on questions of "moral" praise (i.e., praise in response to prosocial actions), I believe that the effects examined should apply to praise in other domains, like in the achievement domain.

Under many theoretical accounts, moral evaluations – including praise and blame – serve to regulate moral behavior and promote cooperation (Curry et al., 2019; Gray et al., 2012; Haidt, 2007). In addition, praise may function more specifically towards building and maintaining relationships (Anderson et al., 2020). For example, because being praised signals a target's social value, praise works to improve interpersonal commitment (Algoe et al., 2016), group commitment (Eisenberger et al., 1986), and the social reputation of the recipient (Henrich & Gil-White, 2001). Combined with the greater costs associated with blame compared to praise, these relationship-building benefits can also result in wider application of praise than blame (Schein et al., 2020).

When assigning moral praise, people are sensitive to factors that reveal the moral agent's character and intentions (Pizarro et al., 2003; Yudkin et al., 2019). People attempt to infer the moral agent's motivations and underlying goals in order to determine whether the person will be likely to cooperate in the future. Even exceptionally generous acts are not seen as praiseworthy if they are viewed as motivated by self-interest (Barasch et al., 2014; Berman et al., 2015; Heyman et al., 2014). When inferring a moral agent's motivations and character, people often use information beyond just the moral act itself. For example, observers make more

positive judgments of someone who engages in a prosocial behavior with a positive expression (e.g., smile) than a negative expression (e.g., frown), inferring that the expression reveals endorsement or condemnation of the behavior itself (Ames & Johar, 2009). Moral agents receive more praise when they make a prosocial decision quickly compared to making the same decision more slowly (Critcher et al., 2013), suggesting that people treat decision speed as revealing the individual's underlying disposition (Morewedge et al., 2014).

Important for the present research, the amount of effort involved in a moral behavior can also influence the praiseworthiness of the behavior (Bigman & Tamir, 2016) as effort is assumed to reflect the relative importance of the goal to the agent (Austin & Vancouver, 1996; Henrich, 2009). Similarly, people make more positive judgments of prosocial behavior that is extraordinary (Futamura, 2018), exceeding expectations of social obligation based on relationship status (McManus et al., 2020), and rare as well as costly (Kraft-Todd & Rand, 2019). Based on this review, I propose that, as a broad category, praise triggers in response to supererogatory actions, prosocial behaviors that exceed norms and expectations.

Supererogatory Behavior

Philosophers (and psychologists in turn) have distinguished between different conceptualizations of morality, focusing on norms surrounding negative and antisocial behavior and on norms surrounding positive and prosocial behavior (Wiltermuth et al., 2010). For example, Fuller (1969) distinguished between a morality of duty – encompassing the minimal criteria for acceptable behavior – and a morality of aspiration – encompassing the virtues that people should, but are not obligated to,

pursue. In a morality of duty, people who violate minimal standards receive sanctions and condemnation, but those exceeding minimal standards do not receive praise (Hamilton et al., 1988). In a morality of aspiration, people can receive praise by exceeding typical standards of virtue. This distinction is consistent with Immanuel Kant's conceptualization (1785/1993) of perfect duties – those that are blameworthy if not fulfilled (e.g., caring for one's child) – and imperfect duties – those that are praiseworthy if fulfilled but are not obligatory (e.g., donating a kidney to a stranger). Consistent with this framing, research has highlighted how adults frequently think of helping behavior as good but not obligatory (Dahl et al., 2020).

A related concept in philosophy and theology – originating in the Roman Catholic Church – is that of supererogation (Flescher, 1994; Urmson, 1958). Supererogation refers to conduct that is morally good but not strictly required – that is, morally desirable action that exceeds one's typical duties, responsibilities, and obligations. What actions count as "supererogatory" depend on the norms of a situation and the roles assigned to/assumed by the moral agent. For example, consider the action of saving another person's life. By itself, this action would likely be considered morally good. However, the extent to which that action is considered supererogatory likely depends on *who* performed the action. Consider the typical duties given to a physician and an engineer. Given the norms surrounding what being a physician entails, saving a life typically falls within the duties for a physician and would therefore be unlikely to qualify as supererogatory by other people (morally beneficial though the action may be). However, because engineers lack similar norms of duty as a function of their profession, saving a life would be more likely seen as

supererogatory for an engineer than a physician. This framing echoes research showing that people process positive behaviors depending on their assessment of the statistical norms surrounding those behaviors (Ngo et al., 2015). This suggests that judgments for positive actions (e.g., praise) depends on how common those actions appear to be, and actions that exceed an agent's typical duties and responsibilities are likely seen as less common.

Moral Judgments as Social Signals

People treat moral judgments, like praise and condemnation, as providing information about both the person making the judgment and the broader normative implications of such judgments. Individuals who engage in condemnation of immoral acts may be perceived as communicating information about their future behavior (Baumeister et al., 2004). For example, observers treat condemnation of immorality as signaling both repudiation of the action by the condemner but also that the condemner possesses good moral character and would behave morally (Jordan et al., 2017; Jordan, Hoffman, Bloom, et al., 2016). Even 7-year-old children treat condemnation as a positive signal of the condemner's character; they infer that the condemner is less likely to transgress in that domain (e.g., someone who condemns a thief is less likely to steal themselves) and judge the condemner as more deserving of punishment if they transgress (Hok et al., 2020). In addition, emotional reactions to moral events (e.g., guilt for accidents) can provide important information about the agent's character and future behavior (Frank, 1988; Prinz, 2004). Thus, moral judgments and emotions can provide information not just on the acceptability or desirability of the behavior but on what the behavior reveals.

In achievement contexts, praise can signal expectations and norms, and is often interpreted as revealing something about the target of the praise. For example, praising a child for being "smart" may actually promote unethical behavior, where the child then cheats in order to maintain that "smart" reputation (Zhao et al., 2017). Even hearing *another* child praised for being "smart" can promote cheating behavior in 5-year-old children (Zhao et al., 2020). Furthermore, when praise is overly positive and enthusiastic, children can develop lowered self-esteem because the praise may be interpreted as expectations being set too high for the child (Brummelman et al., 2017). Praise and encouragement can also indicate that the praise-giver had relatively low expectations for the agent to begin with (Chestnut & Markman, 2018).

Two recent studies have explored praise in the moral domain and found that moral praise can likewise communicate information regarding expectations. Praising a child's self (e.g., "you are such a good helper") appears to increase helping behavior compared to praising a child's action (e.g., "you did a good job helping"; Bryan et al., 2014). However, praising a child for being a "good helper" can hinder children's subsequent prosociality if the child encounters obstacles that cause them to fail at their goal of helping (Foster-Hanson et al., 2020). Thus, in the moral domain, praise can have a potentially powerful impact in shaping moral development and future moral behavior because of what praise communicates regarding expectations. To better understand how best to motivate prosocial action, research needs to address open questions regarding the full range of inferences that people draw from seeing an action praised.

Overview

Across six studies, I examined two foundational questions regarding the psychology of moral praise. First, in Study 1, I aimed to identify the role of duties and actions that exceed duties (i.e., supererogatory actions) in the application of moral praise. I hypothesized that people are more likely to judge an action is praiseworthy when it is supererogatory as compared to when it merely fulfills one's duties. In principle, this need not be the case. People may consider actions and agents praiseworthy simply because those targets bring about some benefit. One alternative prediction would then be that actions that bring about more positive outcomes would be seen as more praiseworthy. However, past research has demonstrated that judgments of positive acts (e.g., praise) are not particularly attuned to the magnitude of the beneficial outcome (Gneezy & Epley, 2014; Klein & Epley, 2014; Yudkin et al., 2019). Because of these findings, I instead focused on duties. I also investigated related questions regarding the connection between how praiseworthy an action is and judgments of blameworthiness, obligations, and likelihood of occurrence.

Second, in Studies 2-5, I address what inferences people make when they see an action or agent receive praise. Specifically, I investigate whether one potential byproduct of moral praise is not simply to indicate that an action is morally good but also that that the action is relatively uncommon and that people are not required to do that. Studies 2-5 build on Study 1 – if supererogatory actions are judged as more praiseworthy than obligatory (i.e., duty-bound) actions, then perhaps people will infer supererogation and obligation based on the presence or absence of praise for an action. Praise may function to communicate and reinforce moral norms, not only through direct reward-learning but instead through the inferences drawn from the act of

praising. All materials, data, analysis codes, and preregistrations can be found at https://osf.io/bf5cw/?view_only=6807cfcd88cc421483bc2288228b6bfc.

Study 1

In Study 1, I first investigated people's lay intuitions regarding the connection between praise and norms regarding duties and expectations. Participants first listed two behaviors, one they considered as falling within a person's duties and one they considered as exceeding a person's duties. I hypothesized that people would be more likely to rate behaviors that exceed duties (i.e., were supererogatory) as more praiseworthy than behaviors that are duties.

In addition to my primary research question, I also conducted two exploratory sets of analyses to more fully contextualize how people think about moral praise and what it may mean. First, I examined the associations between people's judgments regarding their personal likelihood of performing the action, whether people should do the action, praise for doing the action, and blame for not doing the action. One possibility is that the praiseworthiness for doing an action, compared to the blameworthiness for not doing an action, is less connected to people's own assessments of their personal likelihood for doing the action and their judgments of whether people should do the action. Second, I coded what participants listed as actions to examine whether the types of actions listed as duties and as exceeding duties differed in meaningful ways. I predicted that participants would be more likely to mention social targets (i.e., another person) for actions that exceed duties compared to actions that are duties. In addition, I further predicted that if participants listed a

social target, those targets would be more socially distant (e.g., a stranger) for actions that exceed duties than actions that are duties.

Method

Participants

I recruited 150 U.S. participants (M_{age} = 36.30; 47 women, 101 men, 2 did not disclose) from Amazon's Mechanical Turk using CloudResearch's prime panels (Chandler et al., 2019) and provided them with monetary compensation. Most participants (~83%) self-identified their ethnicity as "White". I did not include any exclusion criteria.

Procedure

After giving consent, participants were presented with introductory text about the study, describing how "some acts are seen as a duty" while "other acts are seen as going beyond duty and obligation" (for full text, see OSF link). I then asked participants to list an action that they considered a duty and an action that they considered above and beyond duty. After participants listed the two actions, I asked participants to make four ratings (presented in random order) about each action (each action was on a separate page, presented in random order). Participants made judgments about the likelihood that they would do the action if they were in the appropriate situation (*1 Not at all likely* to *7 Extremely likely*) and whether people should do the action if they were in the appropriate situation (*1 Not at all* to *7 Very much*). Participants also made two evaluative judgments for each action (both from *1 None at all* to *7 A great deal*). Participants reported how much praise someone would deserve if they were in the appropriate situation and did the action. Participants also

reported how much blame someone would deserve if they were in the appropriate situation and did not do the action.

To investigate potential differences in what actions people listed as duties and as above and beyond duty, I had three research assistants code the actions participants listed based on two criteria. First, they coded whether participants mentioned a social target as the beneficiary of the action (coded as 1, e.g., "helping a stranger") or not (coded as 0, e.g., "picking up litter on the ground"; ICC_{duty} = .85, ICC_{above} = .40). The coders were instructed to code self-directed actions (e.g., "taking care of yourself) as being non-social. Second, if participants listed a social target, they coded those targets based on their relative social proximity to the agent (ICC_{duty} = .53, ICC_{above} = .90). Social targets were categorized as either extremely close (coded as 3, e.g., immediate family and similar close others; "helping your parents"), somewhat close (coded as 2, e.g., friends, acquaintances, and neighbors; "helping my neighbor bring in groceries"), or distant (coded as 1, e.g., strangers and generic others; "giving to the needy"). Disagreements between the three coders were resolved by a fourth research assistant, provided with the same set of instructions and the ratings from the other three coders.

Results

Primary Analyses

I first examined whether there were differences between participants' ratings for duties and actions that exceed duty (see Table 1). Participants consistently judged the two types of actions differently. Relative to duties, participants said actions that exceeded duties were behaviors they were less likely to do, people have less obligation to do so (i.e., lower Should ratings), deserve greater praise for doing so, and deserve less blame for failing to do so, *p*s < .001. Framed another way, people think duties are actions that they are more likely to do, people should do more, and people are more

blameworthy for failing to do them, whereas supererogatory actions deserve more praise.

Table 1
Study 1 Results

	Duties	Above and Beyond	
Likely to do the action	6.52 (.88)	5.11 (1.73)	$t(148) = 8.92$, $p < .001$, $d = .73$
People should do the action	6.35 (1.11)	5.26 (1.57)	$t(144) = 8.19$, $p < .001$, $d = .68$
Praise for doing	3.84 (1.96)	5.53 (1.58)	$t(148) = 9.25$, $p < .001$, $d = .76$
Blame for not doing	4.81 (2.01)	3.26 (1.93)	$t(148) = 7.40$, $p < .001$, $d = .60$

Note. Means, SDs, and paired sample *t*-tests comparing ratings for actions that are duties and actions that are above and beyond duty.

Correlations

I next examined the correlations between people's judgments regarding duties and actions that exceed duty (see Table 2). When evaluating both types of actions, participants' ratings of their own likelihood of performing the action positively correlated with their ratings of how much people should do the action, *p*s < .001. Praise judgments did not consistently correlate with the other judgments for either type of action – praise was only significantly correlated with blame for not doing the action when the action was a duty, $p = .003$. On the other hand, blame for not doing the action correlated with likelihood judgments and should judgments for both types of actions, *p*s < .03.

Table 2

Study 1 Correlations

	Actions that are duties				Actions that are above and beyond duties			
	Should	Praise	Blame	Praise vs. Blame	Should	Praise	Blame	Praise vs. Blame
Likelihood	.56***	-.005	.18*	$Z = 1.81$, $p = .07$	.59***	.12	.48***	$Z = 3.51$, $p < .001$
Should		-.08	.22**	$Z = 2.87$, $p = .004$		.15	.40***	$Z = 2.37$, $p = .02$
Praise			.24**				.13	

Note. Correlations between different judgments and Fisher *Z* test for whether judgments of likelihood and should were more closely correlated with praise or blame judgments.
* $p < .05$, ** $p < .01$, ** $p < .001$

As an exploratory set of analyses, I tested whether participants' judgments of likelihood of doing the action and how much people should do the action were more closely correlated with praise for doing or blame for not doing each type of action (see Table 2). Compared to judgments of praise, participants' ratings of likelihood were more correlated with blame for not doing the action, nonsignificant but marginally for duties, $p = .07$, and significantly for above and beyond duties, $p < .001$. Similarly, for both duties and above and beyond duties, participants' judgments of whether people should do the action were more closely correlated with how much blame people deserve for not doing the action than how much praise people deserve for doing the action, ps $< .02$.

Discussion

Consistent with my hypothesis, these results provide evidence that people view praise as more appropriate for actions that exceed a person's duties compared to actions that are a person's duties. That is, the judged praiseworthiness of an action is sensitive to the norms surrounding that action, specifically whether someone is obligated and expected to perform the action or whether performing the action would be outside of their typical obligations and expectations. In addition, the correlational analyses suggest that people see praise as less (if at all) connected with the likelihood of doing the action and whether people should do the action, while blame is more connected with those considerations. These results echo past research that has found that judgments of praise and blame are psychologically distinct and orthogonal to each other (Wiltermuth et al., 2010). One implication from this study is in thinking about what general class of actions receive praise – that of supererogatory behaviors. In

addition, the text analysis provides a clue for what sorts of actions people think of as supererogatory – prosocial behaviors directed towards strangers (McManus et al., 2020).

Study 2

For the remaining studies, I transitioned to understanding people's judgments regarding moral praise and its function and signaling value. Study 1 indicated people more strongly affirm the praiseworthiness of an action when it is supererogatory (i.e., being above and beyond the agent's duties). Essentially, people view praise as more appropriate for normatively exceptional actions than for duty-driven actions. With the remaining studies, I focus on the reverse set of inferences, examining the judgments people make when presented with an action that receives praise. If people are presented with a praised action, do they think that the action is more supererogatory?

With Study 2, I specifically focused on the inferences that people make of the person giving praise. Observers may interpret praise as providing information about the praise-giver's beliefs and motivations (e.g., their expectations, their values, and their goals). I hypothesized that observers would infer not just that the praise-giver thought the praised action was morally right, but also that the praise-giver considered the action relatively rare and that the praise-giver would be more interested in being friends with the moral agent.

Method

Participants

I recruited 500 participants from Prolific.co, an online data collection service (Palan & Schitter, 2018). Sixteen participants failed the attention check, leaving a final sample of 484 (M_{age} = 31.57; 252 women, 221 men, 11 other).

Procedure

Participants were randomly assigned to one of two conditions. All participants read a vignette describing a conversation between two coworkers and what they did the previous weekend. One coworker, Alex, described doing a prosocial action (e.g., "he went to volunteer at a local homeless shelter". The particular action was drawn from a bank of five possible actions (see the OSF link for the full list of actions). The response of the other coworker, Jeremy, then varied by condition. In the *praise* condition, participants read that Jeremy praised Alex and said "Wow, that was really good of you." In the *no praise* condition, participants read that Jeremy simply said "That's cool. I did some chores around the house this weekend."

Participants then completed eight items regarding what they read in the vignette (presented in random order). Five of these questions measured participants' inferences regarding Jeremy and his thoughts. Participants rated their agreement with the following statements (*1 Strongly disagree* to *7 Strongly agree*): (1) Jeremy thinks that what Alex did is morally right. (2) Jeremy thinks that what Alex did is expected of people in the same situation. (3) Jeremy thinks that what Alex did in that situation is relatively rare. (4) Jeremy thinks that Alex is a good person. (5) Jeremy would want to be friends with Alex.

Participants also completed three questions of their more general impressions of what Alex did. One question assessed how common they thought Alex's behavior

was ("Given that situation, how common is what Alex did?" from *1 Very uncommon* to *7 Very common*). One question assessed their thoughts of whether Alex's behavior was effortful and costly ("What Alex did in that situation requires a lot of effort and cost." from *1 Strongly disagree* to *7 Strongly agree*). Participants then reported their judgment of the overall benefit of Alex's behavior ("What Alex did benefited other people." from *1 Strongly disagree* to 7 *Strongly agree*). Finally, participants completed demographic questions and an attention check measure where they were asked to list "purple" as the color of the sky. Participants who did not list purple were excluded from analyses.

Results and Discussion

There was no significant interaction between condition and which prosocial action Alex performed, *p*s > .21, so the effect of action was dropped from subsequent analyses.

Consistent with my predictions, I found that participants made clear inferences about Jeremy and his thoughts when he praised Alex than when he did not praise Alex (Table 3). Specifically, participants in the *praise* condition indicated more strongly that Jeremy thought Alex's behavior was morally good, that Jeremy thought of the behavior as less expected and rarer, that Jeremy thought of Alex as a good person, and that Jeremy would want to be friends with Alex, *p*s < .01. Together, these results strongly indicate that third-party observers make inferences into the expectations, goals, and evaluations of someone based on whether they praise someone for their behavior. However, there were no significant differences between the *praise* and *no praise* conditions regarding participants' judgments of how common Alex's behavior

was, $p = .42$, whether Alex's behavior involved effort or cost, $p = .69$, or whether what Alex did benefited people, $p = .90$.

Table 3
Study 2 Results

	Praise	*No Praise*	
Morally right	6.25 (0.79)	5.59 (1.05)	$t(480) = 7.80$, $p < .001$, d = .71
Expected	3.98 (1.43)	4.38 (1.33)	$t(480) = 3.25$, $p = .001$, d = .30
Rare	4.82 (1.16)	3.80 (1.41)	$t(479) = 8.65$, $p < .001$, d = .79
Good person	6.07 (0.76)	5.41 (0.96)	$t(479) = 8.34$, $p < .001$, d = .76
Friends	5.44 (0.96)	5.16 (1.12)	$t(475) = 2.93$, $p = .004$, d = .27
Common	6.85 (1.38)	6.75 (1.47)	$t(480) = 0.81$, $p = .42$, d = .07
Effort and cost	3.84 (1.52)	3.90 (1.59)	$t(480) = 0.40$, $p = .69$, d = .04
Benefited	6.40 (0.71)	6.39 (0.84)	$t(480) = 0.13$, $p = .90$, d = .01

Note. Means, SDs, and independent sample *t*-tests for the main measures of Study 2.

Study 2 provides consistent evidence that people make inferences based on whether an action receives moral praise. However, I found significant results only regarding participants' evaluations of the person giving the praise, and not on participants' evaluations of the behavior in general. My initial predictions were that participants' would make more general, norm-based inferences based on an action receiving moral praise. However, in retrospect, I realized that the wordings of these

more general questions were potentially ambiguous as to whether they were about the prosocial action that Alex did and was relating to Jeremy, or the conversation itself. In Studies 3-5, I more systematically examine whether people interpret praise as providing information about the descriptive norms for an action.

Study 3

Study 2 demonstrated that people infer mental states – like expectations, values, and goals – about a person who praises someone's behavior. Yet it remains unclear whether people also infer normative information from praise – does praise indicate that a behavior is relatively rare and unique given a particular situation? In Study 3, I used a situation where there is an established prescriptive norm of what people should do – the practice of tithing (i.e., donating 10% of one's income) in Christianity and other religions. To the extent that someone identifies with their church and the church endorses the practice of tithing, the prescriptive norm would be to tithe. However, tithing is generally understood to be a voluntary practice (Dahl & Ransom, 1999; James & Jones, 2011). Therefore, if a person does tithe, people may consider this an exceptional act, exceeding their obligations. I predicted that observers would interpret praise from a church leader to a member for a tithe donation as indicating that such behavior is relatively less common, compared to thanks from the church leader (Study 3a and 3b) or simple acknowledgement of the action (3b). I also investigated the robustness of the effect of praise on norm estimates by using different wordings for communicating the initial norm for tithing (Study 3a).

Study 3a

Method

Participants. I recruited 801 participants from Prolific.co. Thirty-two participants failed the attention check, leaving a final sample of 769 (M_{age} = 33.24; 341 women, 411 men, 17 other).

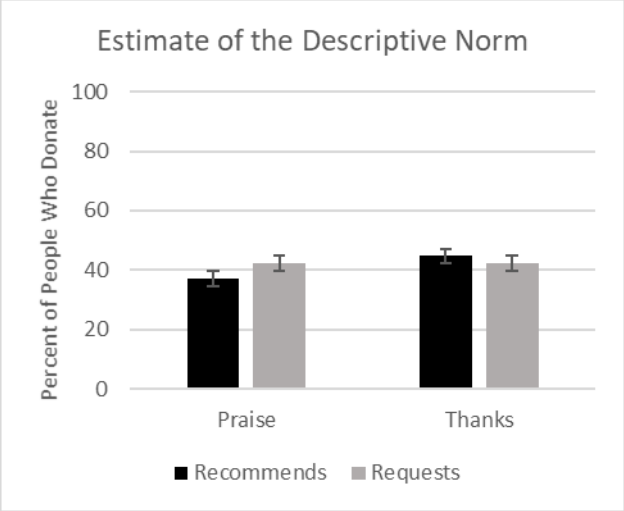
Procedure. Participants were randomly assigned to one of four conditions, based on a 2 (norm: recommends, requests) X 2 (response: praise, thanks) between-subjects design. All participants read a vignette about a man who attends and donates to his church. In the *recommends* condition, participants read that the church "recommends, but does not require" attendees donate 10% of their yearly income to the church. In the *requests* condition, participants instead read that the church "strongly requests" that attendees donate 10% of their yearly income to the church. Participants then learned that John donated 10% of his income to the church, and the church leader called him. In the *praise* condition, the church leader said, "Wow John, that was really good of you." In the *thanks* condition, the church leader said, "John, thank you very much for your donation." Participants then completed an estimate regarding the descriptive norm, indicating what percentage of people in John's church they thought donated at least 10% of their income to the church (from 0% to 100%). Participants then completed demographic questions and the same attention check measure as Study 2.

Results and Discussion

When John's behavior received praise, compared to just gratitude and thanks, participants estimated that significantly *fewer* people donated in the same situation (see Figure 1), $F(1,764) = 6.36$, $p = .01$, $\eta_p^2 = .008$. There was no significant main effect of the particular way in which the norm was framed, $F(1,764) = 1.59$, $p = .21$,

η_p^2 = .002, and no significant interaction, $F(1,764) = 2.42$, $p = .12$, η_p^2 = .003. Praise indicated to observers what was typical in the situation, such that praise made an action seem less common than gratitude. In this particular design, the expression of thanks and gratitude may have been seen as fairly strong, whereby the church leader was still expressing a strong indication of positivity and approval. As such, the effect size of Study 3a may actually be an underestimation of the true effect comparing praise and gratitude. I replicate and extend on this result in Study 3b, which incorporates a different expression of gratitude, an additional comparison condition, and additional measures to further examine observers' judgments.

Figure 1
Study 3a Results



Note. Figure displays means and standard errors for each condition.

Study 3b

Method

Participants. I recruited 600 participants from Prolific.co. Nineteen participants failed the attention check, leaving a final sample of 581 (M_{age} = 30.83; 292 women, 274 men, 15 other).

Procedure. After providing consent, participants read a scenario involving a man, Kevin, who donates to his church. Participants read "At the beginning of the year, in accordance with the tradition of tithing and giving to church, Kevin donated for 10% of his yearly salary to his church." Participants then read about the church leader calling Kevin, with the church leader's response varying by condition between-subjects. In the *praise* condition, the church leader says, "Kevin, wow – that was really good of you." In the *thanks* condition, the church leader says, "Kevin, thank you for your donation." In the *control* condition, the church leader says, "Kevin, we have received your donation."

Participants then completed four measures (presented in random order between participants). Participants indicated their estimation of the descriptive norm by indicating the percentage of people in Kevin's church they thought donated at least 10% of their income to the church (from 0% to 100%). Participants also indicated what percentage of their yearly salary people in Kevin's church donate to the church (from 0% to 100%). Participants completed two additional measures asking, in Kevin's church, whether people are required to donate to the church (from *1 = Definitely not required* to *7 = Definitely required*) and whether people are expected to donate to the church (from *1 = Definitely not expected* to *7 = Definitely expected*). Participants completed the same demographics questions and attention check measure as Study 2.

Results and Discussion

For each measure, I first calculated the omnibus test for an overall difference in conditions and then conducted preregistered contrasts comparing each condition to the others (see Table 4). As hypothesized, there was a significant or marginally significant effect of condition for each measure. Comparing the *praise* condition to the other conditions, I found that participants took praise for an action to indicate that the action was relatively less frequent, less required of people, and less expected of people. In terms of the participants' estimate of the average donation, both praise and thanks for Kevin's donation were interpreted to mean that the average donation was relatively less than if there was mere acknowledgment of the donation.

Table 4
Study 3b Results

	Control	Thanks	Praise	$F(2, 574)$	p	η_p^2
Percent who donate	48.43[a] (29.19)	45.39[a] (28.48)	29.23[b] (25.18)	26.96	< .001	.09
Average donation amount	13.25[a] (16.13)	10.25[b] (10.70)	10.20[b] (13.71)	2.70	.07	.01
Required	3.89[a] (1.95)	3.74[a] (1.96)	3.15[b] (1.69)	7.68	.001	.03
Expected	5.58[a] (1.35)	5.27[b] (1.48)	4.65[c] (1.66)	17.70	< .001	.06

Note. Table displays means and SDs for each condition for the four primary measures, along with omnibus inferential statistics. Superscripts indicate which condition means are significantly different from each other in preregistered contrasts for each measure, at $p < .05$.

Together, these results demonstrate that people interpret praise as indicating both the descriptive norm (i.e., the percentage of people who donate) and the prescriptive norm (i.e., what is required and expected of people) of a situation. In this

way, praise may inadvertently communicate to others that the action was *less necessary* than if the beneficiary had expressed a simple thanks or acknowledgement for the action. Unexpectedly, observers also inferred that acknowledgment for an action was taken to indicate that people tended to be *more generous* in the situation, relative to praise or thanks. Although this effect should be interpreted with caution, this suggests that the *absence* of praise or gratitude may also provide normative information to observers.

Study 4

With Study 4, I aimed to extend the results from Studies 2 and 3 by examining what information praise communicates to observers when the norms are unknown. In the absence of prior experience about how people tend to behave in a particular situation, would observers infer normative information when an action is praised? For example, social and moral norms frequently vary across cultures (Knafo et al., 2009; Miller et al., 1990; Miller & Bersoff, 1992; Shweder et al., 1987; Yuki et al., 2005). When in a new culture, one's prior knowledge about how often people help strangers may not be accurate (Levine et al., 2001). I predicted that in a novel cultural environment, moral praise would act as an indication that the praised behavior was relatively uncommon. I also investigated potential self-other differences in the effect of praise: would participants make different inferences between being the recipient of praise or from seeing another person receiving praise?

Method

Participants

I recruited 1202 participants from Prolific.co. Fifty-six people failed the attention check and were excluded, leaving a final sample of 1146 (M_{age} = 32.03; 595 women, 526 men, 25 other).

Procedure

Participants were randomly assigned to one of six conditions, based on a 2 (perspective: self, other) X 3 (response: praise, acknowledgement, control) between-subjects design. All participants were asked to imagine that they were visiting a foreign country and had only a basic understanding of local customs and practices. Specifically, participants were told they were not "quite sure which actions are expected and required of people". Participants then read that a couple of days into the trip, they were traveling to a coffeehouse and while crossing the street they see an elderly man struggling to walk and carrying heavy bags. In the *self* condition, participants read that they run over to the man, help him cross the street, and walk him to his building. In the *other* condition, participants read that they see a woman run over to the man, help him cross the street, and walk him to his building. After participants read about the elderly man being helped home, they then read that they continued to the coffeehouse and ordered a drink. Participants read that the barista was at the window and could see what happened outside. While the barista is preparing the drink, she converses with the participant. In the *praise* condition, the barista praises either the participant (*self* condition) or the woman (*other* condition) for helping the elderly man, saying "That was a *really* good thing that [you/she] did." In the *acknowledgment* condition, the barista acknowledges that the participant (*self* condition) or the woman (*other* condition) helped the man but then only gives additional information about the man, saying "He has lived here for a long time." In

the *control* condition, the barista does not acknowledge or comment on what happened with the elderly man, instead saying "Nice day we're having, right?"

Participants then completed three measures (presented in random order). Participants estimated what percentage of people in this country they thought would help an elderly person in need (from 0% to 100%). Participants completed two additional measures asking, in this country, how required is it that people help the elderly (from *1 = Definitely not required* to *7 = Definitely required*) and how expected is it that people help the elderly (from *1 = Definitely not expected* to *7 = Definitely expected*). Participants finally completed demographic items and the same attention check measure as Study 2.

Results and Discussion

To assess the influence of praise on observer's estimates of the descriptive and injunctive norms, I conducted a 3 (response: praise, acknowledgment, control) X 2 (perspective: self, other) general linear model on each of the primary measures (See Table 5 for descriptive statistics).

For estimates of the percentage of people in the country who would help, I found a main effect of perspective, $F(1,1140) = 21.76$, $p < .001$, $\eta_p^2 = .02$, and a main effect of response, $F(2,1140) = 16.22$, $p < .001$, $\eta_p^2 = .03$. These main effects were qualified by a significant interaction, $F(2,1140) = 4.10$, $p = .02$, $\eta_p^2 = .01$. When imagining themselves as the moral agent, participants gave lower estimates for how many people in the country who would help after receiving either *praise* or *acknowledgement*, compared to no mention of the action by the bystander in the *control* condition. When imagining another person as the moral agent, participants gave lower estimates when the bystander praised the action, but gave roughly equal estimates when the bystander simply acknowledged the action or did not comment on it at all. I did not predict this particular interaction a priori, so I caution against reading

too much into its significance. As speculation, this may suggest that people treat even acknowledgment of their own moral actions as praise, or at least information that the action is less common in the society. Whereas, when considering another person's moral action, observers must be more explicitly praising in terms of providing information about the statistical norms. At the least, observers interpreted praise for an action, compared to no mention of the action, to indicate that fewer people in the society performed that action.

Table 5
Study 4 Results

		Praise	Acknowledgment	Control
Percentage estimate of people in this country who would help	Self	48.94 (23.48)	48.85 (23.98)	56.49 (21.43)
	Other	50.89 (23.04)	60.02 (21.35)	62.01 (21.84)
How required is it that people help the elderly?	Self	2.99 (1.66)	3.35 (1.65)	3.81 (1.61)
	Other	3.27 (1.75)	3.63 (1.73)	3.74 (1.75)
How expected is it that people help the elderly?	Self	3.73 (1.65)	4.08 (1.57)	4.56 (1.58)
	Other	4.09 (1.69)	4.65 (1.59)	4.95 (1.50)

Note. Means and standard deviations for each condition for the three primary measures.

For judgments of whether helping is required in the country, I found a significant main effect of response, $F(2, 1138) = 13.61$, $p < .001$, $\eta_p^2 = .02$. There was no significant effect of perspective, $F(1,1138) = 2.66$, $p = .10$, $\eta_p^2 = .002$, and no significant interaction, $F(2, 1138) = 1.30$, $p = .27$, $\eta_p^2 = .002$. Compared to

acknowledgment or no comment, praise for a moral action was taken to indicate that the action was less required of people in the country.

For judgments of whether helping is expected in the country, I found a significant main effect of perspective: $F(1,1138) = 21.49$, $p < .001$, $\eta_p^2 = .02$, whereby participants judged that the action was less required in the *self* condition than in the *other* condition. This makes sense, given that participants in the *other* condition may have interpreted the moral agent as a local, and therefore falling in the category of people "in this country". Crucially for the hypotheses, I found a significant main effect of response, $F(1, 1138) = 26.48$, $p < .001$, $\eta_p^2 = .04$, such that praise for an action, compared to simple acknowledgement or no mention, was interpreted to mean that the action was less expected of people in the country. There was no significant interaction, $F(1, 1138) = 0.46$, $p = .63$, $\eta_p^2 = .001$.

These results build on Studies 2-3, providing additional evidence that people treat moral praise as providing a signal regarding the norms of a culture for a specific behavior. When a person does not know what the particular norms are for a situation, people infer from moral praise that an action is relatively uncommon and is not required or expected of people. I also found evidence that such inferences apply to both one's own actions and the actions of other people.

Study 5

With Study 5, I examined people's beliefs regarding moral praise as a learning mechanism. If an agent receives praise for their action, what inferences do people make regarding that agent's future behavior? I compared the inferences people make for praise and for blame, comparing whether these moral judgments are seen as having

differential effects in shaping future behavior. Recent research has highlighted that reward and punishment (and praise and blame by extension) are organized around communicative principles, whereby evaluative feedback is used to signal target behaviors (Ho et al., 2019; Sarin et al., 2021).

Integrating the results from Studies 1-4, praise is interpreted to mean that the praised behavior is not necessarily required nor expected of the agent. Observers may therefore predict that the agent is less likely to repeat the praised behavior, or at least not perform an action of greater prosociality (e.g., having been praised for donating $4, the agent does not then donate $5). Praise is a positive evaluation, so the agent can simply do what they did again in order to maintain the positive evaluation. Praise may therefore be seen as unlikely to change an agent's behavior. On the other hand, blame indicates that an action was *not desired* by the evaluator. Therefore, observers may predict that the agent *will* change their behavior because of any blame they have received. Even young children recognize that people who are punished are less likely to transgress than those who are not punished (Bregant et al., 2016).

In Study 5, I examine the predictions observers make of an agent receiving either praise or blame (compared to a control neutral statement) for a donation behavior in an experimental economic game, similar to the dictator game (e.g., Bardsley, 2008; Bolton et al., 1998). Across conditions, participants were presented with the same behavior, where an agent is endowed with money (e.g., $10) and then distributes a slight majority of money to themselves (e.g., $6) and the remainder (e.g., $4) to another person. The moral agent then receives from a third-party observer either praise for this action, condemnation for this action, or no evaluation. Participants then

made inferences about the agent's future behavior and the norms of the situation. I decided on the particular distribution to navigate two potential trade-offs. First, instead of keeping all of the endowed money, the agent is giving at least *some* money to the other person, which could conceivably be considered generous and thus praiseworthy. Second, because the distribution is not an even distribution, and therefore inequitable, the action could still be considered blameworthy. It seems unlikely that participants would put much credence in blame if the agent had given a majority of the money to the other person.

Method

Participants

I recruited 506 participants from Prolific.co and provided monetary compensation. I excluded 65 participants for failing at least one of the comprehension checks, and 1 participant for not answering any of the dependent measures. The final sample included 440 participants (M_{age} = 32.29; 236 women, 194 men, 10 other).

Procedure

After providing consent, participants are introduced to an experimental economic game with three players. Participants were told that the game is being conducted by a university research team and that the players in the game would operate under conditions of anonymity. Each player is assigned a different role in the game, acting as either the Sender, the Receiver, and the Observer. The Sender receives $10 and then makes a decision for how to divide that money between themselves and the Receiver. The Receiver simply receives however much money they were given by the Sender. The Observer learns about the decision made by the Sender and can

provide text-based feedback to them. After reading about the rules of the game, participants were asked three comprehension check questions: 1) “How much money is given to the Sender to start with?”, 2) “Can the Receiver make any decisions in the game?”, and 3) “What does the Observer do?” Participants had to correctly answer all three questions to be included in the analyses; participants that incorrectly answered any of these questions were automatically sent to the end of the study.

Participants were then randomly assigned to one of three between-subjects conditions. In all conditions, participants learned that in the first round of the game, the Sender decided to give $4 to the Receiver and keep $6 for themselves, and then the Observer sent a message to the Sender. In the *praise* condition, the Observer expressed approval of the action and said, “That was really nice of you. If I was the Receiver, I'd be pretty happy.” In the *blame* condition, the Observer expressed condemnation of the action and said, “That wasn’t very cool of you. If I was the Receiver, I’d be pretty mad.” In the *control* condition, the Observer did not express any evaluation of the action and simply said, “First round finished.”

Participants then completed four questions, presented in randomized order. All questions were answered using a slider scale from $0 to $10 in $0.10 intervals. One question served as the measure of participants’ estimates of behavioral change. Specifically, participants were asked to predict how much money the same Sender would give in a second round of the game, with another $10 but new people as the Receiver and the Observer. One question served as the measure of participants’ estimate of the descriptive norm, asking participants to estimate how much money they think people in the position of the Sender generally give to the Receiver (i.e.,

what is the average amount of money given). One question served as a measure of participants' estimates of the injunctive norms in the situation - I asked participants to indicate how much money they thought Receivers expect to get from Senders. Finally, I asked participants to indicate their own personal beliefs about the situation, asking them how much money they thought Senders should give to Receivers. After answering these four questions, participants completed a short demographics survey.

Results and Discussion

Per my registration, I conducted an omnibus test to detect overall differences between conditions, followed up with planned contrasts comparing each of the conditions to the other conditions (see Table 6).

Table 6
Study 5 Results

	Praise	Blame	Control	$F(2, 437)$	p	η_p^2
Second round estimation	4.00[a] (1.57)	4.49[b] (1.44)	3.97[a] (1.26)	6.26	.002	.03
Average amount sent	3.40[a] (1.66)	3.93[b] (1.35)	3.83[b] (1.36)	5.57	.004	.02
Amount that Receivers expect	3.58[a] (1.91)	4.63[b] (1.54)	4.33[b] (1.85)	13.75	< .001	.06
Amount that Senders should give	4.61[a] (1.28)	4.87[a] (1.18)	4.68[a] (1.39)	1.49	.23	.007

Note. Table displays means and SDs for each condition for the four primary measures, along with omnibus inferential statistics. Superscripts indicate which condition means are significantly different from each other in preregistered contrasts for each measure, at $p < .05$.

Consistent with my prediction that the nature of the evaluative feedback would influence judgments of future behavior, I found a significant effect of condition on

predictions of donation amount in a second round, $p = .002$. Specifically, I found that when the Sender received blame and condemnation from the Observer (compared to the *praise* and *control* conditions), participants estimated that the Sender would give more money to a new Receiver in another round of the game. Given an initial behavior that was below an even distribution of the money (i.e., $5), participants anticipated that third-party blame and condemnation would prompt others to be *more* generous in the future. However, participants thought that praise would on average produce no change in behavior.

There was a significant main effect of condition on estimates of the average amount of money given by Senders, $p = .004$. Specifically, when the Sender was given third-party praise by the Observer for a $4.00 donation (compared to the *blame* and *control* conditions), participants estimated that Senders on average give relatively less money. Consistent with the previous studies, people interpret praise for an action as indicating that the action is relatively more prosocial than what people do on average (i.e., compared to the descriptive norm). Interestingly, participants in the *blame* and *control* conditions estimated that the average behavior was not significantly different from the Sender's behavior of sending $4.00, $ps > .13$. Together, this pattern of results indicates that, when an action receives praise, the implied expectation is that people are on average less generous. When an action receives blame, the implied expectation is *not* that people are on average more generous: instead, participants estimated no significant difference between the statistical norm and the agent's behavior.

Similarly, there was a significant difference between conditions on estimates on the amount of money that Receivers generally expect to receive, $p < .001$. When

the $4.00 donation was praised (compared to the *blame* and *control* conditions), participants judged that Receivers tended to expect relatively less money. One possible interpretation of this finding is that participants interpreted the Observer's comments as a reaction for how that person would have felt in the position of the Receiver. Thus, participants would have taken the Observer and their reaction as a potential substitute for how Receivers in general would think and feel.

For participants' personal beliefs of how much Senders should give, I found no overall significant difference between conditions and no differences from the planned contrasts, *p*s > .23.

General Discussion

The present studies aimed to address two open questions regarding the nature of moral praise. The first question I addressed was whether one potential dimension by which praiseworthy actions become praiseworthy is whether the action exceeded the agent's duties and responsibilities. In Study 1, I found that participants on average judged actions that exceed a person's duties, compared to a person's duties, as being more praiseworthy. On the other hand, participants judged that failing to meet duties as more blameworthy, and that people are more likely to fulfill and should fulfill duties. This suggests that certain actions are judged as praiseworthy because of their supererogatory nature, which has implications for judgments of how often such behaviors occur and whether people are obligated to act in that way.

The second question I addressed was what inferences people draw from seeing an act receive moral praise. Supporting the hypotheses, I found that people infer information both about the person giving the praise and about the norms surrounding

the action. Building upon its theorized functions (Anderson et al., 2020; Schein et al., 2020), this work highlights that praise may additionally serve as a social signal, communicating the praise-giver's expectations and desires (Study 2) and the norms of whether certain prosocial actions are expected and required of people (Studies 3-5).

Implications

I believe these findings have important implications in understanding how and what people learn from social and moral judgments like praise. Specifically, my work demonstrates that people infer that praise does not simply provide information about the value or desirability of an action (i.e., whether some action is *good*). Instead, praise provides information about the surrounding normative context for the action: how statistically common that action is and whether people are required or expected to do that action. The presence or absence of praise may thus function to communicate information about what the baseline or reference behavior is, and such reference points can have important implications for how people construe their social environment (Chestnut & Markman, 2018). By seeing attention drawn to certain behaviors through praise, observers may thus learn about what is and is not expected and morally required of people.

Another implication of this research concerns people's inferences for *why* a behavior is appraised as uncommon. If praise connotes that a behavior is relatively rare, observers may subsequently wonder why the behavior is rare. This implication may apply more for achievement contexts, but one potential inference that people may make is that the praised action, although valued, is relatively difficult or costly for people (or at least, that specific agent) to do. For example, if two people both donate

$100 but only one person receives praise, one inference that people may make is that the praised agent has less money than the non-praised agent. Praise may thus communicate not just what the broad norm is but specifically what are the duties and expectations of the agent themselves.

While the present studies have focused on praise for moral actions, I believe that these findings can speak to praise for other types of actions. There is a large literature on praise for competence-based achievements (e.g., academics, athletics, artistic skills), especially in the context of motivation and learning (e.g., Henderlong & Lepper, 2002; Mueller & Dweck, 1998). While research on praise for moral behaviors and for achievement behaviors frequently operate in isolation, I believe that these findings would apply equally well for the latter behaviors. For example, praising someone for their artistic accomplishment likely communicates similar information about norms and expectations regarding that behavior: that such achievements are relatively uncommon and not necessarily expected of the agent.

Future Directions

The present studies provide an initial examination of the interplay between moral norms and moral praise. Of course, additional work is necessary to replicate and extend these findings to provide a more complete understanding of the inferences people make from praise. One promising avenue for future research involves factors regarding both the praise-giver and the praise-receiver. The impact of moral praise may depend on *who* is giving it. For example, past research has shown that group leaders can have an especially powerful impact on group norms (Crandall et al., 2018; Freeman et al., 2004; Georgeac et al., 2019; Lemoine et al., 2019; Munger, 2017;

Padilla et al., 2007; Sims & Brinkman, 2002). Thus, praise from a group leader (versus just an average group member or an outgroup member) may not only be especially reinforcing for the recipient but also provide a clear signal of what are the group's norms and values. Similarly, the perceived moral character of the praise-giver likely matters, whereby people with higher perceived character are likely trusted in their praise more than those with lower perceived character. On the praise-recipient side, future work should examine how praise may differ when given to a novice versus an expert. For example, praising a child for a behavior likely provides a weaker signal for the norms than praising an adult for the same behavior, or at least the inferences to be drawn from the praise may be more restricted (e.g., how expected it is that a young child would do this behavior).

Another open question is how praise relates to other socially-oriented positive evaluations, like gratitude. Do observers make similar judgments regarding whether an action is common or required if it receives gratitude as compared to praise? As part of this distinction, one potentially important component is the relationship between the moral agent and the moral commentator. For example, feelings of gratitude are most likely to be felt by beneficiaries of a moral agent's behavior, which may complicate the signal value of the gratitude – is the gratitude simply in response to receiving help or from the exceptional nature of the help? Beneficiaries can of course also praise moral agents, but observers may conflate praise from beneficiaries (e.g., "you're a good person") with gratitude from beneficiaries (e.g., "thank you for what you did"). In a pilot study (N=99) conducted on Prolific.co, I asked participants to describe what they would say if they were to praise a person for doing something morally good. I

found that 37% of participants included mention of thanking that individual, suggesting that people frequently think of "praise" and "gratitude" in similar ways. Future work should more directly examine similarities and differences between expressions of praise, expressions of gratitude, and other positive evaluations (e.g., liking).

Additional work should also explore the differences in how people make judgments of (and from) moral praise and moral blame (Anderson et al., 2020). For example, one underexplored question is how praise and blame relate to each other in terms of being public expressions versus private attitudes. Blame seems to lend itself better to being a private attitude (in the sense of holding someone causally/morally responsible), whereas praise seems to lend itself more naturally to being a communicative expression. To praise someone typically *is* to publicly acknowledge their good qualities, whereas to blame someone need not be public but could instead be a privately held attitude (Malle et al., 2014; Shaver, 1985; Weiner, 1995). One potential consequence of this asymmetry is that there should be a greater match between someone cognitively judging a target as praiseworthy and then publicly praising that target than someone cognitively judging a target as blameworthy and then publicly blaming that target. Perhaps the more direct negative analogue to praise is condemnation and punishment, which may have a similarly public, communicative function (Ho et al., 2019; Sarin et al., 2021).

Conclusion

Moral praise is a rich social judgment that both responds to and reflects normative considerations. Beyond just having a role in reinforcement learning, one of

praise's functions may lie in its communicative ability regarding expectations and desires. While these studies offer an initial exploration regarding the inferences people draw from moral praise, future research should extend these findings to further understand the relation between moral praise and moral norms.

CHAPTER 3

"FALSE POSITIVE" EMOTIONS, RESPONSIBILITY, AND MORAL CHARACTER[1]

"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.[2] 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 et al., 2014; Shaver, 1985; Weiner, 1995), and an agent is more likely to

[1] Published as: Anderson, R. A., Kamtekar, R., Nichols, S., & Pizarro, D. A. (2021). "False positive" emotions, responsibility, and moral character. *Cognition*, *214*, 104770. © Elsevier B. V. This paper is not the copy of record and may not exactly replicate the authoritative document published in the journal. Please do not copy or cite without author's permission. The final article is available at: https://doi.org/10.1016/j.cognition.2021.104770

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

be blamed when she intentionally brings about a harmful outcome (Cushman, 2008; Sloman et al., 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 & Shultz, 1990; Darley et al., 1978; Shultz et al., 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" (Badar & Marchuk, 2013; Perkins, 1939; Royzman & Kumar, 2004).

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 (Mandel & Dhami, 2005; Smith et al., 2002; Weiner et al., 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 I 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?

False Positive Emotions

In this project, I 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 (for discussion, see Kamtekar & Nichols, 2019). 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? My goal was to examine how people might use a specific person's display of a false positive moral emotions[3] (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). I 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,

[3] For my 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 I focus on guilt for one's own harms and gratitude for being the recipient of another's beneficence.

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.

Inferring Moral Character

My hypotheses are based 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 et al., 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 et al., 2014) and define personal identity largely in moral terms (Heiphetz et al., 2018; 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 et al., 2013; Fedotova et al., 2011; Gray et al., 2012; Pizarro et al., 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 (Gold & Weiner, 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 et al., 1994).

The Current Studies

I 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 this project, I 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). My 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 the preregistrations, analyses reported here exclude certain participants, although none of the conclusions are substantively altered if these participants are included (see OSF link).

Study 1

Study 1 served as an initial test of the hypothesis, allowing me to examine the judgments that observers make of agents who feel the false positive moral emotions of guilt and gratitude. As a central focus, I wanted to test whether false positive moral emotions would be perceived as reliable predictors of an agent's moral character. Participants saw 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

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)	Character; 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*	Character; 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*	Character; 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

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). I

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.

Method

Participants

I 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 > .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).

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 minutes), while the other half was told that he had plenty of time (the train was leaving in 30 minutes). 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.

Results and Discussion

Coffee Spill Scenario

I 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 ($r_{Spearman\text{-}Brown}$ = .94). I 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 ($r_{Spearman\text{-}Brown}$ = .94). I also combined the three questions measuring Janet's predicted guilt from various moral infractions (α = .94) into a single index of predicted guilt. For summary of results see Table 2.

Consistent with my 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* < .001. Participants also judged that Janet should have displayed significantly more emotion in the *no guilt* condition than in the *guilt* condition, $p < .001$.

Table 2

Study 1 Results for the Guilt scenario

	Guilt *M* (*SD*)	No Guilt *M* (*SD*)	*t*(369)	*p*	*d*
Moral character	5.42 (1.03)	3.01 (1.32)	18.71	<.001	1.94
Social likability	5.03 (1.18)	2.59 (1.50)	17.49	<.001	1.81
Likelihood of future moral offense	2.99 (1.45)	4.81 (1.45)	12.10	<.001	1.26
Likelihood of future guilt and shame	5.90 (1.13)	2.98 (1.59)	20.45	<.001	2.13
Responsibility	4.08 (1.85)	4.36 (1.71)	1.54	.125	0.16
Agent should have displayed more emotion	4.38 (0.73)	5.73 (1.47)	11.27	<.001	1.17
Victim should have displayed more emotion	4.21 (0.78)	4.29 (0.70)	1.03	.30	0.11

Note. Judgments of moral character served as my 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.

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 my 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.

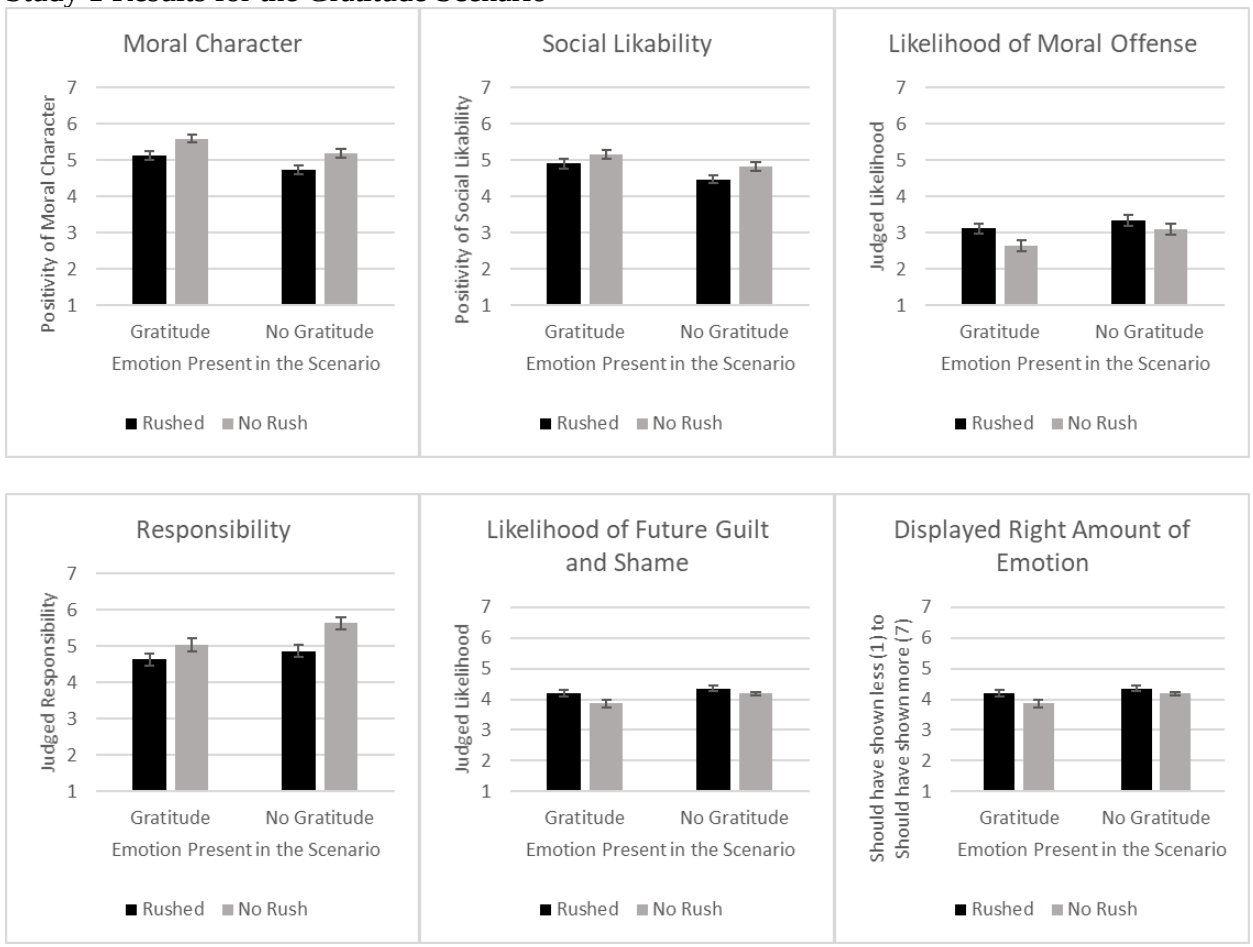
Train Ticket Scenario

As in the Coffee Spill scenario, I calculated a single index of moral character ($r_{Spearman\text{-}Brown}$ = .86), social likability ($r_{Spearman\text{-}Brown}$ = .84), and predicted guilt (α = .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 Figure 1). "Grateful" Peter was rated as having better moral character, $F(1, 357) = 15.33$, $p < .001$, $\eta_p^2 = .04$, and greater social likeability, $F(1, 357) = 11.28$, $p = .001$, $\eta_p^2 = .03$. When Peter expressed high gratitude, he was also judged as less likely to commit a minor moral offense, $F(1, 357) = 5.17$, $p = .02$, $\eta_p^2 = .01$, as less responsible for what happened, $F(1, 357) = 6.08$, $p = .01$, $\eta_p^2 = .02$, as expressing more guilt for moral transgressions, $F(1, 357) = 23.68$, $p < .001$, $\eta_p^2 = .06$, and as not needing to display more gratitude, $F(1, 357) = 6.17$, $p = .01$, $\eta_p^2 = .02$.

Similarly, when Peter was not rushed, he was rated as having better moral character, $F(1, 357) = 19.28$, $p < .001$, $\eta_p^2 = .05$, as being more socially likable, $F(1, 357) = 7.41$, $p = .007$, $\eta_p^2 = .02$, as being less likely to commit a minor moral offense, $F(1, 357) = 5.98$, $p = .02$, $\eta_p^2 = .02$, as more responsible for what happened, $F(1, 357)$

Figure 1

Study 1 Results for the Gratitude Scenario



Note. Graphs display means and standard errors for each condition. Primary measure of interest is moral character.

= 6.08, p = .001, η_p^2 = .03, as expressing more guilt for moral transgressions, F(1, 357) = 3.97, p = .04, η_p^2 = .01, and as not needing to display more gratitude, F(1, 357) = 6.54, p = .01, η_p^2 = .02. Contrary to my initial predictions, there were no significant interactions between expressions of gratitude and whether or not Peter was rushed, all ps > .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 I 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.

Study 2

In Study 2, I aimed to both replicate and extend the findings from Study 1 by making several modifications to the materials and design. Specifically, I 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. I 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.

Study 2 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 towards the self (for a review, see McCullough et al., 2001), gratitude towards 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.

Method

Participants

I recruited 408 U.S. participants through MTurk. My initial aim was to recruit at least 100 participants per condition, which would provide power > .80 for the primary hypotheses based on the observed effect sizes in Study 1. I 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 the preregistration, I 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.

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), I 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." I 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. Participants then read 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)[4], and were asked how

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

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, I asked participants to assign blame for Janet/praise for Peter for what happened in each scenario. I 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, I 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.

Results and Discussion

Coffee Spill Scenario

I 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, $\alpha = .95$) and a composite index for social likability (how much they like Janet and how much they would want to get to know Janet, $r_{Spearman\text{-}Brown} = .93$). I 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, $r_{Spearman\text{-}Brown}$ = .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, $r_{Spearman\text{-}Brown}$ =.92). See Figure 2 for a summary of the results.

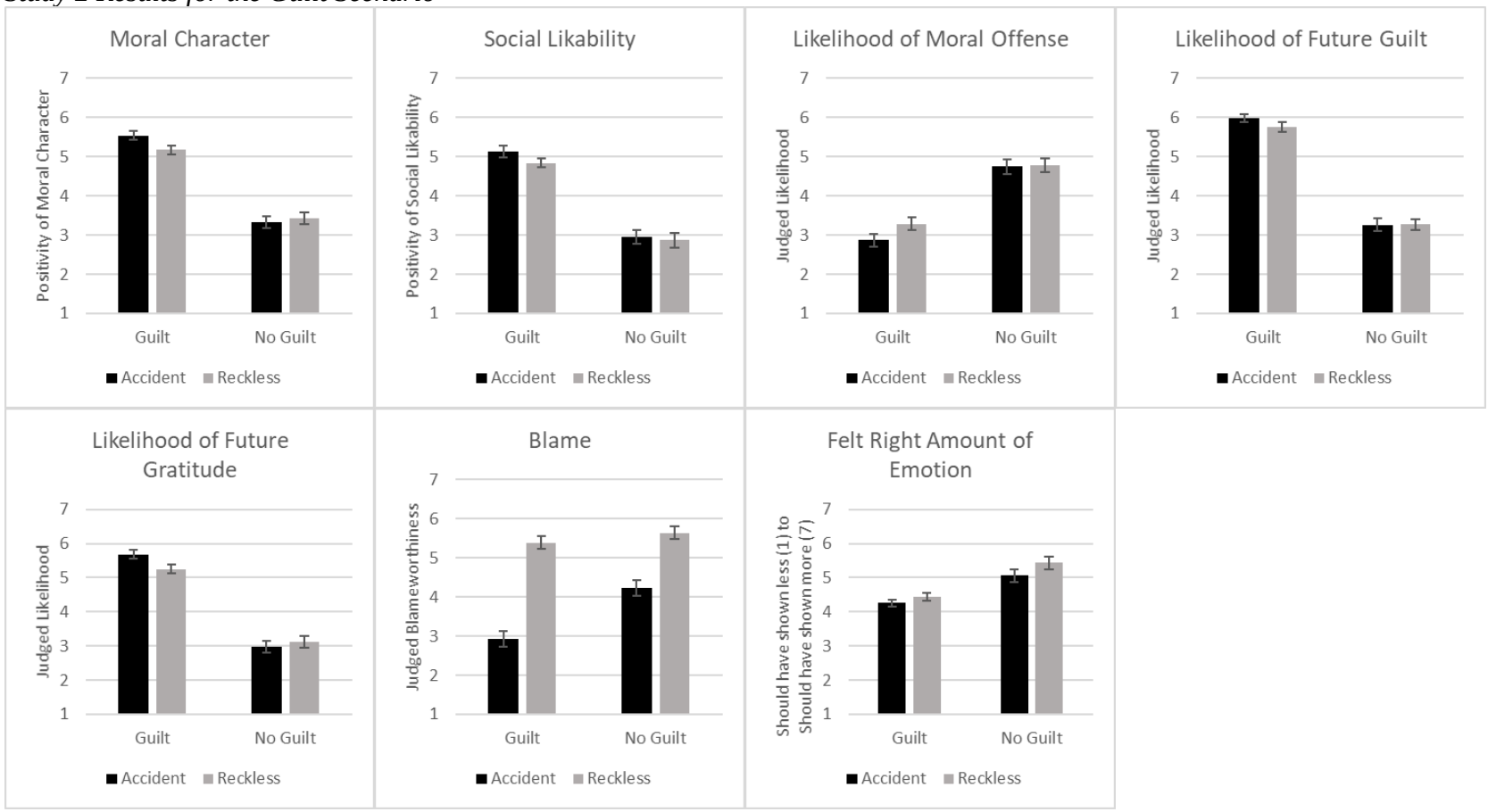
Replicating the finding from Study 1 and consistent with my hypothesis, there was a significant main effect of emotion on judgments of general moral character, such that participants judged Janet as having better character when she felt guilt than when she felt no guilt, $F(1, 303) = 231.45$, $p < .001$, $\eta_p^2 = .43$. There was no significant main effect of responsibility (i.e., accident vs. recklessness conditions) and no interaction between responsibility and guilt, *p*s > .07.

Likewise, participants judged Janet as being more socially likable when she felt guilt compared to when she felt no guilt, $F(1, 303) = 172.86$, $p < .001$, $\eta_p^2 = .36$. There was no significant main effect of responsibility (i.e., accident vs. recklessness conditions) and no interaction between responsibility and guilt, *p*s > .25.

Consistent with my hypotheses, participants judged Janet as more likely to commit a minor moral offense in the *no guilt* condition than in the *guilt* condition, $F(1, 303) = 97.96$, $p < .001$, $\eta_p^2 = .24$. There was no significant main effect of responsibility, and no significant interaction between responsibility and guilt, *p*s > .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, $F(1, 303) = 372.43$, $p < .001$, $\eta_p^2 = .55$. For predictions of guilt, there was no significant main effect of responsibility and no interaction between responsibility and guilt, *p*s > .68.

Figure 2

Study 2 Results for the Guilt Scenario



Note. Graphs display means and standard errors for each condition. Judgments of moral character are the primary measure of interest.

Participants judged that Janet was more likely to feel gratitude in other situations in the *guilt* condition than in the *no guilt* condition, $F(1, 303) = 269.94$, $p < .001$, $\eta_p^2 = .47$. There was no significant main effect of the responsibility condition on such judgments, $F(1, 303) = 0.98$, $p = .32$, $\eta_p^2 = .003$, and no significant emotion by responsibility interaction, $F(1, 303) = 3.70$, $p = .055$, $\eta_p^2 = .01$.

Consistent with my 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, $F(1, 302) = 17.79$, $p < .001$, $\eta_p^2 = .06$. 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, $F(1, 302) = 111.93$, $p < .001$, $\eta_p^2 = .27$. These main effects were qualified by a significant interaction between emotion and responsibility, $F(1, 302) = 8.02$, $p = .005$, $\eta_p^2 = .03$. Breaking down this interaction, participants judged Janet as more blameworthy in the *no guilt* condition than in the *guilt* condition in the *accident* condition, $t(302) = 5.03$, $p < .001$, $d = .58$, but there was no significant difference in blame between the *no guilt* and *guilt* conditions in the *reckless* condition, $t(302) = 0.97$, $p = .33$, $d = .11$.

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, $F(1, 302) = 34.79$, $p < .001$, $\eta_p^2 = .10$. There was no significant effect of responsibility on judgments of how much guilt Janet should have felt, $F(1, 302) = 3.369$, $p = .066$, $\eta_p^2 = .01$. There was no significant interaction between emotion and responsibility, $F(1, 302) = 0.49$, $p = .48$, $\eta_p^2 = .002$.

In summary, the 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.

Train Ticket Scenario

As with the *Coffee Spill* scenario, I combined items to form single measures of general moral character ($\alpha = .94$), social likability ($r_{Spearman\text{-}Brown} = .91$), predicted guilt ($r_{Spearman\text{-}Brown} = .85$), and predicted gratitude ($r_{Spearman\text{-}Brown} = .89$). See Table 3 for a summary of results. Consistent with my 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 an agent's moral character and future behavior. Even if gratitude is directed towards someone fulfilling their duties, observers treat such gratitude as indicative of the agent's character.

Table 3

Study 2 Results for the Gratitude Scenario

	Gratitude *M* (*SD*)	No Gratitude *M* (*SD*)	*t*(305)	*p*	*d*
Moral character	5.67 (0.85)	3.90 (1.29)	14.19	<.001	1.62
Social likability	5.55 (0.93)	3.46 (1.38)	15.84	<.001	1.77
Likelihood of future guilt	5.99 (1.00)	4.25 (1.59)	11.47	<.001	1.31
Likelihood of future gratitude	6.18 (0.82)	4.03 (1.69)	14.09	<.001	1.62
Praise	4.25 (1.79)	2.55 (1.61)	8.76	<.001	1.00
Likelihood of future act of charity	5.93 (1.01)	3.79 (1.48)	14.68	<.001	1.68
Agent should have felt more emotion	4.28 (0.91)	4.60 (1.39)	2.36	.02	0.27

Study 3

Study 3 expands the 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, I 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 (Critcher et al., 2020; Uhlmann et al., 2015). I 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, I 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 I 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 used (see Westfall et al., 2015).

Method

Participants

I recruited 120 U.S. participants (47% female, M_{age} = 37.36) through MTurk. This sample provides power > .95 to detect sample sizes observed in Studies 2. I did not include any comprehension checks or exclusion criteria.

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 minutes 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."

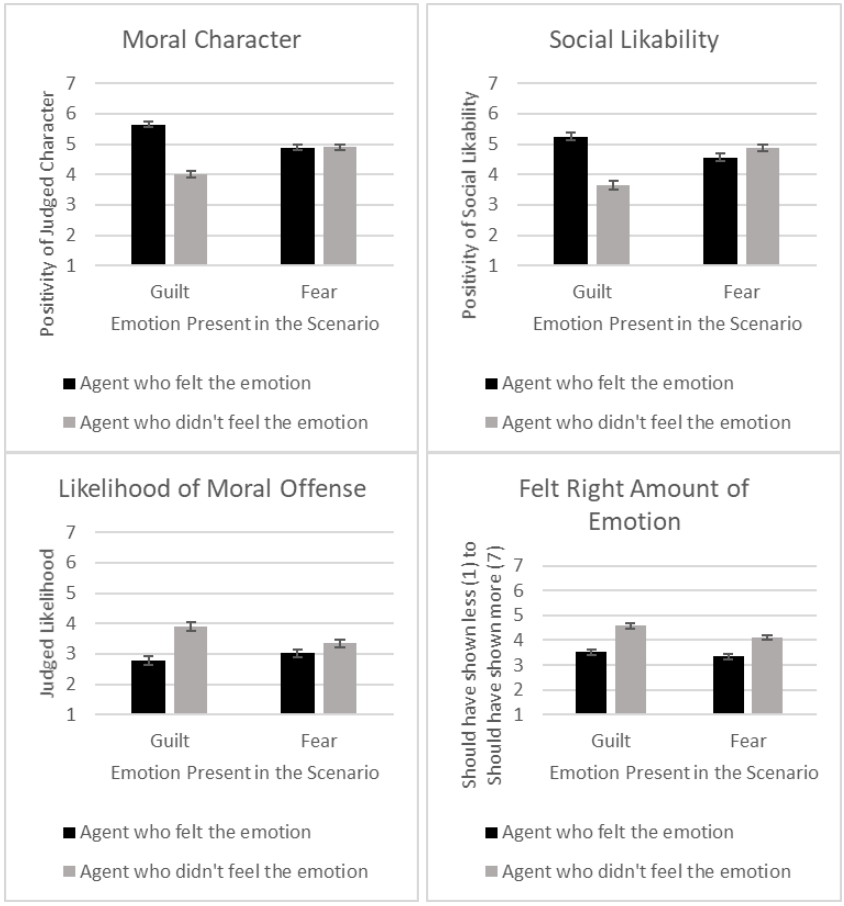
For all four agents across the two scenarios, participants answered the same moral character items (all αs > .82) and social likability items (all $r_{Spearman\text{-}Brown}$s > .82) from Study 2, the likelihood of committing a minor moral offense item from Study 2, and whether the agent felt the right 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*).

Results and Discussion

Consistent with my hypothesis that participants use the presence of certain emotions to inform their judgments of moral character, I found a significant main effect of emotion presence, $F(1, 119) = 124.34$, $p < .001$, $\eta_p^2 = .51$, and the predicted interaction between emotion type and emotion presence on judgments of general moral character, $F(1, 119) = 99.52$, $p < .001$, $\eta_p^2 = .46$. 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 (Figure 3). There was a nonsignificant effect of the emotion type on judgments of moral character, $F(1, 119) = 0.63$, $p = .43$, $\eta_p^2 = .005$.

Figure 3
Study 3 Results



Note. Graphs display means and standard errors for each condition. Judgments of moral character served as the primary test of my hypothesis – it is not simply any false positive emotion that observers use to infer character.

I next examined how participants assessed the agent's social likability. There was a significant main effect of emotion type, $F(1, 119) = 13.84$, $p < .001$, $\eta_p^2 = .10$, a significant main effect of emotion presence, $F(1, 119) = 51.65$, $p < .001$, $\eta_p^2 = .30$, and a significant interaction, $F(1, 119) = 73.17$, $p < .001$, $\eta_p^2 = .38$. 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, $F(1, 115) = 41.86$, $p < .001$, $\eta_p^2 = .27$, and a significant interaction between emotion type and emotion presence on the predicted likelihood of the agent committing a minor moral offense, $F(1, 115) = 11.01$, $p = .001$, $\eta_p^2 = .09$. 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. There was a nonsignificant effect of emotion type, $F(1, 115) = 1.86$, $p = .18$, $\eta_p^2 = .02$.

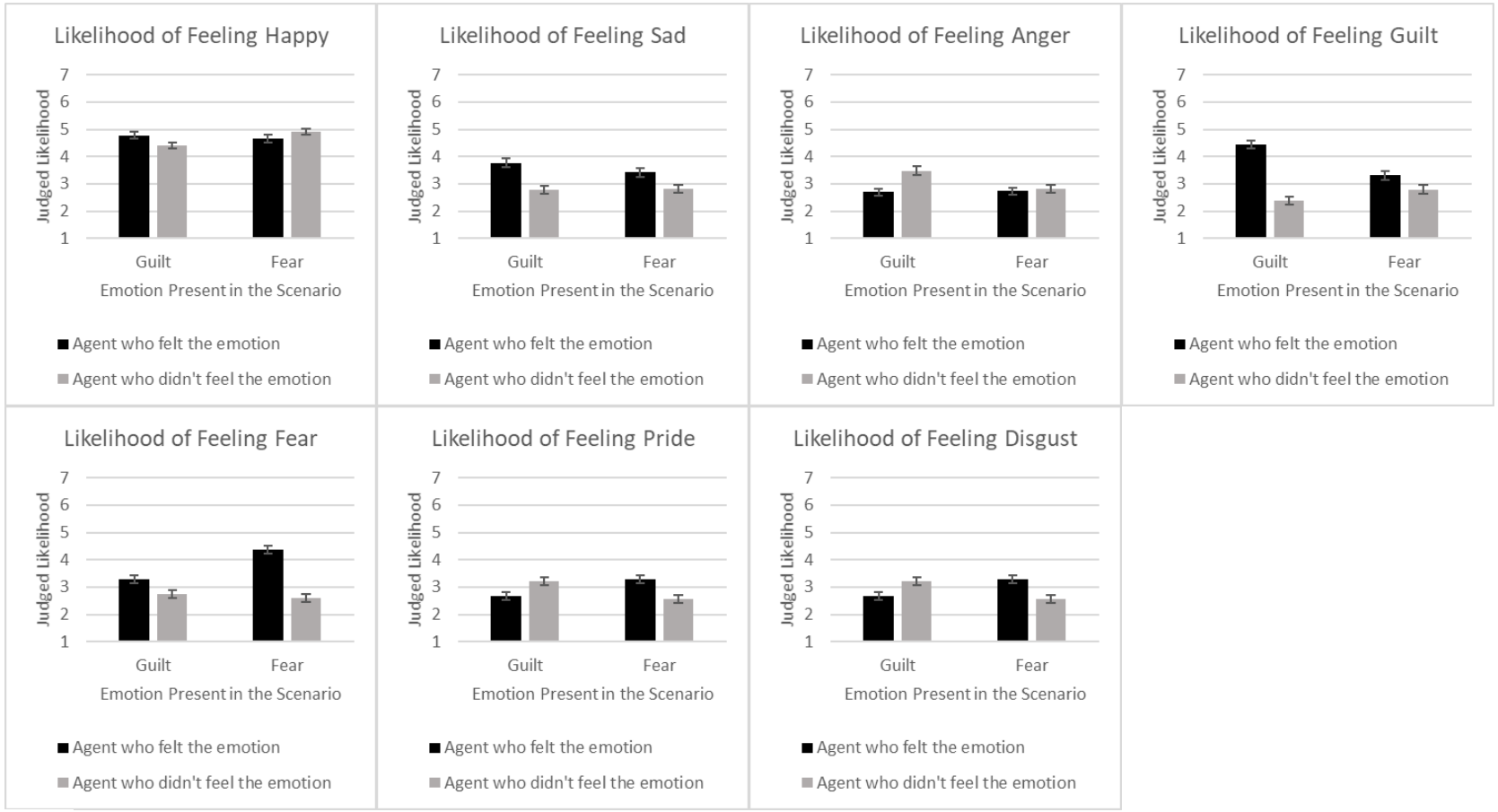
For evaluations of whether the agents felt the right amount of emotion, there was a significant main effect of both emotion type and emotion presence. Participants thought that agents in the fear scenario should have felt relatively less emotion than agents in the guilt scenario, $F(1, 118) = 9.59$, $p = .002$, $\eta_p^2 = .08$. 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, $F(1, 118) = 59.32$, $p < .001$, $\eta_p^2 = .33$. There was no significant

interaction between emotion type and presence, $F(1, 118) = 3.58$, $p = .06$, $\eta_p^2 = .03$. 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), $t(118) = 1.68$, $p = .096$, $d = .13$. 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 = 1.29) and the agent who *did not feel the emotion* ($M = 1.63$, SD = 1.17), $t(119) = 1.15$, $p = .25$, $d = .06$.

To further test the role of different emotions in judgments of moral character, I examined participants' predictions of the likelihood that different agents would experience various emotions (see Figure 5). 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.

Figure 4
Study 3 Emotion Ratings



Note. Mean judged likelihood for each agent from each scenario for a variety of target emotions (error bars indicate SE).

Study 4

In Study 4, I investigated whether the 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 observed were driven by the *false positive* guilt (as predicted) 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.

I 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 (Davis et al., 1995; Kamtekar & Nichols, 2019; Mandel & Dhami, 2005; Niedenthal et al., 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, I 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.

Method

Participants

I recruited 438 U.S. participants through MTurk. I aimed for at least 100 participants per condition to achieve power > .90 based on the observed effects in Studies 1-3. Per the preregistration, I excluded 3 participants for completing the study in less than 30 seconds, leaving a final sample size of 435 (46% female, M_{age} = 36.77).

Design

I 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, I 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."

I then asked participants to complete the same measures from Study 2 (moral character [$\alpha = .88$], social likability [$r_{Spearman\text{-}Brown} = .87$], predicted guilt [$r_{Spearman\text{-}Brown} = .75$], predicted gratitude [$r_{Spearman\text{-}Brown} = .77$], 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, I 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), $r_{Spearman\text{-}Brown}$ = .61, from *1 = not at all* to *7 = very much*).

Results and Discussion

Per the preregistration, I 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 my hypothesis, there was a significant difference between conditions on judgments of moral character, $F(3, 431) = 24.67$, $p < .001$, $\eta_p^2 = .15$. 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, $t(431) = 3.02$, $p = .003$, $d = .39$; and no significant difference between the Janet *guilt* condition and the *vicarious guilt – near* condition, $t(431) = 0.15$, $p = .88$, $d = .02$. 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 the participants did not view Sarah's guilt in this condition as inappropriate.

Table 4
Study 4 Results

	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)

Note. Means and standard deviations for each condition. Ratings were made on a 1-7 scale. Judgments of moral character served as the primary measure of interest.
[a] *Guilt* and *No Guilt* ratings significantly differed, $p < .05$.
[b] *Guilt* and *Vicarious Guilt – Near* ratings significantly differed, $p < .05$.
[c] *Guilt* and *Vicarious Guilt – Far* ratings significantly differed, $p < .05$.

There was a similar pattern in terms of judgments of social likability, with an overall significant difference between conditions, $F(3, 431) = 15.24$, $p < .001$, $\eta_p^2 = .10$. Planned contrasts revealed a significant difference in judgments of likability between the Janet *guilt* condition and the Janet *no guilt condition*, $t(431) = 5.70$, $p < .001$, $d = .81$; a significant difference between the Janet *guilt* condition and the

vicarious guilt – far condition, $t(431) = 4.44$, $p < .001$, $d = .60$; and no significant difference between the Janet *guilt* condition and the *vicarious guilt – near* condition, $t(431) = 0.79$, $p = .43$, $d = .11$.

There was an overall significant difference between conditions on predictions of the agent's likelihood of experiencing guilt in future situations, $F(3, 431) = 40.85$, $p < .001$, $\eta_p^2 = .22$. 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, $t(431) = 1.73$, $p = .09$, $d = .26$, and in the *vicarious guilt – far* condition, $t(431) = 0.69$, $p = .52$, $d = .09$, 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, $F(3, 431) = 30.26$, $p < .001$, $\eta_p^2 = .17$. 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, $t(431) = 6.79$, $p < .001$, $d = .87$; expected Sarah in the *vicarious guilt – near* condition to be more likely to express gratitude in other situations, $t(431) = 2.38$, $p = .02$, $d = .35$; and expected Sarah in the *vicarious guilt – far* condition to be equally likely to experience gratitude in future situations, $t(431) = 0.63$, $p = .53$, $d = .09$.

I next examined how much blame participants assigned to the agent for what happened. Again, there was a significant overall difference between conditions, $F(3, 430) = 25.04$, $p < .001$, $\eta_p^2 = .15$. Compared to the *guilt* condition, participants

assigned significantly more blame to the agent in the *no guilt* condition, $t(430) = 2.83$, $p = .005$, $d = .38$. For vicarious targets, participants assigned significantly less blame in the *vicarious guilt – near* condition, $t(430) = 5.68$, $p < .001$, $d = .85$, and similar amounts of blame in the *vicarious guilt—far* condition ($M = 2.64$, SD = 1.84), $t(430) = 1.17$, $p = .24$, $d = .15$. Consistent with the 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. One 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).

I 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 my predictions, I found a significant overall difference between conditions, $F(3, 426) = 15.88$, $p < .001$, $\eta_p^2 = .10$. Contrasts revealed significantly higher judgments of likelihood in the *no guilt* condition than in the *guilt* condition, $t(426) = 5.61$, $p < .001$, $d = .83$. I found no significant difference in judgments of likelihood between the *guilt* condition and both the *vicarious guilt – near* condition, $t(426) = 0.69$, $p = .49$, $d = .09$, and the *vicarious guilt – far* condition, $t(426) = 1.36$, $p = .17$, $d = .19$. 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.

I also found a significant overall difference between conditions on participants judgments of Janet/Sara's dispositional neuroticism, $F(3, 431) = 75.05$, $p < .001$, $\eta_p^2 = .34$. Compared to the *guilt* condition, participants judged Janet as less neurotic in the *no guilt* condition, $t(431) = 6.20$, $p < .001$, $d = .96$; judged Sarah as more neurotic in the *vicarious guilt – near* condition, $t(431) = 6.81$, $p < .001$, $d = .94$; and judged Sarah as more neurotic in the *vicarious guilt – far* condition, $t(431) = 6.17$, $p < .001$, $d = .81$.

Finally, I found a significant difference of condition on participants' judgments regarding whether the agent felt the "right" amount of guilt in response to the accident, $F(3, 431) = 51.40$, $p < .001$, $\eta_p^2 = .26$. 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, $t(431) = 1.70$, $p = .09$, $d = .30$; 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, $t(431) = 5.74$, $p < .001$, $d = .73$.

Together, these 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.

Study 5

In Study 5, I 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 et al., 1995). I predicted that individuals would be more likely to trust an agent who displayed false positive guilt compared to an agent who did not.

Method

Participants

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

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 the hypotheses[5]. The first situation read "Imagine you are in a crowded

[5] 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

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, participants then read 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 (I 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 one of the other players would be randomly selected 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.

---

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.

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, I asked participants who they would most prefer as a partner in the TG, Player 1 or Player 2. As indicators of trust, I 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%).

Results and Discussion

Consistent with my hypotheses, as well as with the results from the 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, I 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 the hypotheses, participants transferred more money to the *guilty* partner ($M = 30.15$) than the *non-*

guilty partner (M = 14.18), Z = 6.83, p > .001, r = .56. Participants also reported expecting to receive more money back from the *guilty* partner (M = 41.24%) than the *non-guilty* partner (M = 18.43%), Z = 7.61, p < .001, r = .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.

Study 6

In Studies 1-5, I 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, I 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 previously developed and validated. Specifically, I 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). I 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, I believe that this would constitute the first evidence that this tendency to over-experience moral emotions might be a reliable predictor of underlying moral character.

Method

Participants

I 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).

Design

Participants responded to four hypothetical scenarios and a battery of individual difference measures (described below). 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, I 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 et al., 1995), the 20-item Machiavellianism scale (Mach-IV; Christie & Geis, 1970), and the Single Item Narcissism Scale (SINS; Konrath et al., 2014). I 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.

Results and Discussion

Guilt

Participants who reported feeling guilty in the false positive scenario also tended to report feeling guilty in the true positive scenario, $r(204) = .23$, $p = .001$, 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 (α = .86; r[205] = -.19, p = .006), Machiavellianism (α = .69; r[205] = -.16, p = .02), and narcissism (r[204] = .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

Table 5
Study 6 Results

	False Positive Guilt	True Positive Guilt	False Positive Gratitude	True Positive Gratitude
Psychopathy	-.13†	-.19**	-.16*	-.24***
Callous Affect	-.21**	-.22***	-.21**	-.25***
Interpersonal Manipulation	-.09	-.13†	-.13†	-.18*
Erratic Lifestyle	.03	-.08	-.02	-.12†
Machiavellianism	-.13†	-.16*	-.13†	-.11
Narcissism	-.13†	-.20**	-.03	-.15*
No Meaning	.09	-.09	.01	-.22**
Social Desirability	-.01	.05	.04	.06

Note. Correlations between individual difference measures (including the psychopathy subscales) and self-reported guilt and gratitude for true positive and false positive scenarios. I was most interested in the correlations with Psychopathy, particularly the Callous Affect subscale. † < .1, * p < .05, ** p < .01, *** p < .001.

meaninglessness (α = .91; r[205] = -.09, p = .22) and social desirability (r[205] = .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 the primary hypothesis that moral character, as measured by individual

differences in “dark triad” personality traits, is associated with the degree to which a person experiences guilt in both false positive scenarios and true positive scenarios.

Examining the three factors of the psychopathy scale individually, I found that participants who scored higher in callous affect (α = .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 = .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 = .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.

Gratitude

Participants who reported feeling gratitude in the false positive scenario also tended to report feeling gratitude in the true positive scenario, $r(205) = .17$, $p = .02$, 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 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 my 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.

I 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.

General Discussion

Collectively, these 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). I 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, I 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. My 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.

I 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 et al., 2014).

I 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)(Everett et al., 2016), cooperating without carefully calculating costs and benefits (Jordan, Hoffman, Nowak, et al., 2016), and a willingness to engage in third-party punishment (Jordan, Hoffman, Bloom, et al., 2016).

Finally, I found that inferences of moral character from an agent's false positive moral emotions may actually be warranted. In Study 6, I 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).

Moral Emotions as Moral Predictors

These findings make sense given the body of research that has focused on the social function of moral emotions (Algoe & Haidt, 2009; Hutcherson & Gross, 2011; Tangney et al., 1996, 2007). For example, feelings of guilt can motivate attempts to repair a damaged relationship (Schmader & Lickel, 2006; Wicker et al., 1983). Moreover, agents who anticipate feeling aversive emotions like guilt and regret for a decision tend to avoid making that decision (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; 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). These results suggest that false positive emotional responses specifically may provide an additional, and apparently informative, source of evidence for one's propensity towards moral emotions and moral behavior.

My results can also provide insight into understanding collective guilt (i.e., guilt in response to harm done by members of one's ingroup, Wohl et al., 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 et al., 2008).

Limitations and Future Directions

These 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 my findings, as well as to address potential limitations. One such limitation is that the studies utilized several single-item measures, which may have reliability issues (Wanous & Reichers, 1996), so future work should aim to replicate a more robust set of measures. Furthermore, while I used a variety of different vignettes and methods (e.g., the trust game in Study 5), the 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 the 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 et al., 2012).

While I have provided an initial examination of guilt and gratitude, I 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). The 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.

These 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 studies, I 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; for a critique of the usefulness of "moral disgust" as a concept, see Landy & Piazza, 2019). Because the

present 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, I would hypothesize that these effects might generalize to other 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 et al., 2001; McCullough et al., 2013).

One clear limitation of these studies is that my samples were exclusively drawn from U.S. populations using online recruitment methods, limiting the ability to generalize the findings to other populations (especially when it comes to the correlational findings from Study 6). For instance, online convenience samples may differ in important ways from random samples of the nation's general population (Arditte et al., 2016). In addition, researchers have documented differences in norms regarding the experience and expression of emotions across cultures (Mesquita, 2001; Tracy & Robins, 2007; Tsai et al., 2006). Specifically, for the purposes of my hypotheses, there has been research showing cultural variability in the sorts of circumstances that reliably trigger both guilt (Bear et al., 2009; Onwezen et al., 2014; Stipek, 1998) and gratitude (Morgan et al., 2014; Naito et al., 2005). What counts as a "false positive" moral emotion is likely a contextualized judgment that varies 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 et al., 2018; McAuliffe et al., 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.

Conclusion

I 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. I 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, these findings highlight the

importance of emotions and emotional reactions in people's conceptions of what it means to be a "good person."

CHAPTER 4

GENERAL DISCUSSION

Overall, the research described here examines one source – the moral reactions of others – that people use to make sense of their moral environment. That is, one method for understanding the moral world is to see how other's think of and respond to the moral world. First, I demonstrated that moral praise, the judgment of a person's prosocial actions, provides information about both the person giving the praise but also the broader situational norms surrounding the praised action. Second, I provided evidence that false positive expressions of moral emotions are used by observers to predict that agent's moral character and how that agent would behave in true positive situations.

In both cases, people are using someone's moral reaction (e.g., their evaluative judgments and reflexive emotions) to get a sense of the moral landscape without having to directly observe the full extent of that landscape. Praise (and other similar judgments) offers people a summary statement of the broader situational norms surrounding the behavior, without having to necessarily observe multiple instances of that behavior to make a judgment for how frequent or expected that behavior is. Likewise, moral emotions like guilt and gratitude allow observers to get a sense of the target's character and predict how that person may behave in the future without having to see multiple instances of that person's behavior. In a variety of contexts, people are quick to make judgments about each other (Ambady & Rosenthal, 1992; Ballew & Todorov, 2007; Bar et al., 2006; Frank et al., 1993; Todorov et al., 2015; Todorov &

Porter, 2014), and observing a person's moral reactions may serve as one source of information for making such judgments.

A Brief History of Moral Psychology

A simplified, and hopefully not too glib, history of the field of moral psychology is that researchers have focused on the psychology of *act* evaluation, examining how people determine what actions are right or wrong. Beginning with Piaget (1965) continuing with his student Kohlberg (e.g., Kohlberg, 1981; Tapp & Kohlberg, 1971), moral psychology centered on how children reasoned through various moral dilemmas involving potential harms and injustices. Revisions, challenges, and updates to this prevailing trend expanded the conversation to include a variety of moral values (e.g., Gilligan, 1982; Graham et al., 2009; Janoff-Bulman & Carnes, 2013; Rai & Fiske, 2011; Shweder et al., 1987), differentiating between moral rules and conventional rules (Smetana, 1981), evaluations of intentional versus accidental harms (Darley et al., 1978), and the role of emotions and affective processes in moral judgment (e.g., Greene, 2001; Haidt, 2001; Pizarro, 2000; Valdesolo & DeSteno, 2006). Many of these experimental and theoretical paradigms again were developed to understand how people make judgments regarding the morality of particular actions: a person did Behavior *X* – Is Behavior *X* acceptable or not?

More recently, there has been a growing appreciation for the role of the moral *agent* in understanding moral psychology (e.g., Critcher et al., 2020; Goodwin et al., 2014; Gray et al., 2012; Hartley et al., 2016; Heiphetz et al., 2018; Strohminger & Nichols, 2014; Uhlmann et al., 2015). Moral actions are done by *people,* and those

people have beliefs, intentions, and desires that inform their actions. From a very young age, humans attend to other social agents and attempt to understand, and therefore predict, their goals and behaviors (Brooks & Meltzoff, 2005; Helming et al., 2014; Kinzler et al., 2007; Onishi & Baillargeon, 2005; Woodward, 1998). To fully understand moral psychology requires an appreciation for how people make judgments about moral agents and their mental states (Helzer & Critcher, 2018). This appreciation then leads to two important insights. First, people make inferences not simply about a moral action but about the person deemed responsible or connected to that action (e.g., Chakroff et al., 2017; Critcher et al., 2013; Everett et al., 2016). Second, people make judgments about a moral action based on who is responsible for that action (e.g., Gray & Wegner, 2011; Masicampo et al., 2014; Siegel et al., 2017). Updating our experimental design: Person *A*, with Mental State *K*, does Behavior *X* – Is Person *A* a good person? Given *A* and *K*, was Behavior *X* acceptable? Will Person *A* do Behavior *Y* in the future?

Even further than an appreciation of moral agents, I argue that an additional component has been missing, or at least underacknowledged, from the field: the informative role of judgments, reactions, and responses observed in others. It is clear that people attend to both agents and their actions in their own moral assessments, but what counts as an "action" should include both overt behaviors and also judgments (e.g., praise and condemnation; Hok et al., 2020; Jordan et al., 2016, 2017) and emotional responses (e.g., guilt, gratitude, and happiness; Ames & Johar, 2009). Such responses and judgments can have a powerful impact on social norms (Crandall et al., 2018; Munger, 2017; Paluck & Shepherd, 2012), or at least perceptions of what the

norms are (Tankard & Paluck, 2017). Exposure to such responses may be at the root of how moral values develop in the first place (Graham et al., 2011; Haidt, 2007; Nichols et al., 2016; Shweder et al., 1987). In this way, we can think of a more interconnected understanding of moral psychology: Agents perform actions, who are judged by observers, with such judgments (if public) then evaluated by secondary observers, with both sets of observers going on to be agents in their own right, and so forth. The experimental design grows more complex: Person *A*, with Mental State *K*, does Behavior *X*; Person *B*, with Values *W*, then has Reaction *J* – Will Person *B* do Behavior *X*? Will Person *A* do Behavior *X* again, having seen Reaction *J*? Will Person *C*, seeing Reaction *J*, do Behavior *X* or hold Values *W*? Given *J*, is Behavior *X* consistent with Values *W*?

Future Directions

For each of the individual projects, I discussed potential future directions. I believe most, if not all, of them can be applied to the broader research topic regarding moral reactions, especially questions of who is having and receiving the moral reaction (which may be particularly informative for identifying group memberships and affiliations). Here I will discuss two additional avenues for future research.

One avenue for potential research involves questions surrounding accuracy and the "rationality" of the inferences people make from others' moral reactions. That is, to what degree do these reactions reflect an accurate state of the world, and how accurate are the inferences that people make from observing these reactions? As I demonstrated in Chapter 2, people make inferences about both descriptive and injunctive norms from seeing an action receive praise. Are they justified in doing so,

such that praised behaviors *are* actually less common and less required of people than nonpraised behaviors? One potential answer is that such reactions are, at least in part, driven by a rational learning process, whereby individuals gain an understanding of moral norms based on statistical inference (Nichols et al., 2016). In Chapter 3, I found evidence that people who reported experiencing guilt tended to score lower on subclinical measures of psychopathy (especially the callous affect subscale), suggesting that observers may be justified in using such expressions in evaluating moral character. Other work has shown that third-party punishment of unfair behavior is a reliable signal of the punisher's trustworthiness (Jordan, Hoffman, Bloom, et al., 2016). However, these are perhaps unique situations and I recognize that more work can and should be done to understand the relation between how moral reactions reflect moral behavior (both at the norm level and at the agent level).

A second avenue for potential research can examine the connections between different moral reactions (e.g., a judgment like blame and an emotion like anger). Do people make similar inferences from seeing similarly-valenced reactions for the same event? For example, past theorizing has attempted to disentangle such conflation of terms (e.g., blame is not the same thing as anger; for a review, see Malle et al., 2014). In addition, in Study 3 of Chapter 2 I found that participants treated praise for an action and gratitude for that action as signaling different norm levels. However, it is possible that participants perceived a difference in overall positivity between the praise and the gratitude – if the response were more evenly matched in positivity (e.g., "That was a very good thing you did" vs. "We are very grateful for what you did"), would perceivers still infer different norms? Additional research should explore how,

when, and why observers may make different inferences based on the particular responses for a moral action.

Conclusion

There are many ways in which people attempt to understand their moral world. Here, I have argued for one particular method: observing the moral reactions of others. First, I have shown that moral praise, which is responsive to the supererogatory nature of an action, can signal to others that the action was supererogatory. Second, I have shown that false positive expressions of moral emotions (like guilt and gratitude) are treated as signals of an agent's character and future behavior. As a practical consideration, people should be aware of what potential signals their reactions may be sending both to the intended and to their unintended recipients.

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

Algoe, S. B., Kurtz, L. E., & Hilaire, N. M. (2016). Putting the "you" in "thank you": Examining other-praising behavior as the active relational ingredient in expressed gratitude. *Social Psychological and Personality Science, 7*(7), 658–666. https://doi.org/10.1177/1948550616651681

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

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256

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

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Sciences*, *24*(9), 694–703. https://doi.org/10.1016/j.tics.2020.06.008

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*(4), 1241. https://doi.org/10.2307/1127807

Arpaly, N., & Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *93*(2), 161–188. JSTOR.

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, *120*(3), 338–375. https://doi.org/10.1037/0033-2909.120.3.338

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, *117*(5), 2332–2337. https://doi.org/10.1073/pnas.1911517117

Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

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), 1–48. https://doi.org/10.1007/s10609-012-9187-z

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*(46), 17948–17953. https://doi.org/10.1073/pnas.0705435104

Bandura, A. (1977). *Social learning theory.* (pp. viii, 247). Prentice-Hall.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278. https://doi.org/10.1037/1528-3542.6.2.269

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

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, *11*(2), 122–133. https://doi.org/10.1007/s10683-007-9172-2

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

Baumeister, R. F., Zhang, L., & Vohs, K. D. (2004). Gossip as cultural learning. *Review of General Psychology*, *8*(2), 111–121. https://doi.org/10.1037/1089-2680.8.2.111

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25–37. https://doi.org/10.1016/j.cognition.2016.10.024

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

Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The braggart's dilemma: On the social rewards and penalties of advertising prosocial behavior. *Journal of Marketing Research*, *52*(1), 90–104. https://doi.org/10.1509/jmr.14.0002

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*(12), 1654–1669. https://doi.org/10.1037/xge0000230

Bolton, G. E., Katok, E., & Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory*, *27*(2), 269–299. https://doi.org/10.1007/s001820050072

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

Bregant, J., Shaw, A., & Kinzler, K. D. (2016). Intuitive jurisprudence: Early reasoning about the functions of punishment. *Journal of Empirical Legal Studies*, *13*(4), 693–717. https://doi.org/10.1111/jels.12130

Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, *22*(6), 723–742. https://doi.org/10.1037/0012-1649.22.6.723

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*(6), 535–543. https://doi.org/10.1111/j.1467-7687.2005.00445.x

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

Brummelman, E., Nelemans, S. A., Thomaes, S., & Orobio de Castro, B. (2017). When parents' praise inflates, children's self-esteem deflates. *Child Development*, *88*(6), 1799–1809. https://doi.org/10.1111/cdev.12936

Bryan, C. J., Master, A., & Walton, G. M. (2014). "Helping" versus "being a helper": Invoking the self to increase helping in young children. *Child Development*, *85*(5), 1836–1842. https://doi.org/10.1111/cdev.12244

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

Chakroff, A., Russell, P. S., Piazza, J., & Young, L. (2017). From impure to harmful: Asymmetric expectations about immoral agents. *Journal of Experimental Social Psychology*, *69*, 201–209. https://doi.org/10.1016/j.jesp.2016.08.001

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*(5), 2022–2038. https://doi.org/10.3758/s13428-019-01273-7

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

Chestnut, E. K., & Markman, E. M. (2018). "Girls are as good as boys at math" implies that boys are probably better: A study of expressions of gender equality. *Cognitive Science*, *42*(7), 2229–2249. https://doi.org/10.1111/cogs.12637

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

Cialdini, R. B. (2003). Crafting Normative Messages to Protect the Environment. *Current Directions in Psychological Science*, *12*(4), 105–109. https://doi.org/10.1111/1467-8721.01242

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276. https://doi.org/10.1016/0010-0277(89)90023-1

Crandall, C. S., Miller, J. M., & White, M. H. (2018). Changing norms following the 2016 U.S. presidential election: The Trump effect on prejudice. *Social Psychological and Personality Science*, *9*(2), 186–192. https://doi.org/10.1177/1948550617750735

Critcher, C. R., Helzer, E. G., & Tannenbaum, D. (2020). Moral character evaluation: Testing another's moral-cognitive machinery. *Journal of Experimental Social Psychology*, *87*, 103906. https://doi.org/10.1016/j.jesp.2019.103906

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

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69. https://doi.org/10.1086/701478

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

Dahl, A., Gross, R. L., & Siefert, C. (2020). Young children's judgments and reasoning about prosocial acts: Impermissible, suberogatory, obligatory, or supererogatory? *Cognitive Development*, *55*, 100908. https://doi.org/10.1016/j.cogdev.2020.100908

Dahl, G. B., & Ransom, M. R. (1999). Does where you stand depend on where you sit? Tithing donations and self-serving beliefs. *American Economic Review*, *89*(4), 703–727. https://doi.org/10.1257/aer.89.4.703

Darley, J M, & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*(1), 525–556. https://doi.org/10.1146/annurev.ps.41.020190.002521

Darley, John 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. https://doi.org/10.2307/1128594

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

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., & Lee, J. J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, *23*(12), 1549–1556. https://doi.org/10.1177/0956797612448793

Dijk, C., de Jong, P. J., & Peters, M. L. (2009). The remedial value of blushing in the context of transgressions and mishaps. *Emotion*, *9*(2), 287–291. https://doi.org/10.1037/a0015081

Dijk, C., Koenig, B., Ketelaar, T., & de Jong, P. J. (2011). Saved by the blush: Being trusted despite defecting. *Emotion*, *11*(2), 313–319. https://doi.org/10.1037/a0022774

Eisenberger, R., Huntington, R., Hutchison, S., & Sowa, D. (1986). Perceived organizational support. *Journal of Applied Psychology*, *71*(3), 500–507. https://doi.org/10.1037/0021-9010.71.3.500

Eriksson, K., Strimling, P., & Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, *129*, 59–69. https://doi.org/10.1016/j.obhdp.2014.09.011

Everett, J. A. C., 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*(3), 316–317. https://doi.org/10.1177/1754073911402387

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67–73. https://doi.org/10.1177/0963721417738825

Foster-Hanson, E., Cimpian, A., Leshin, R. A., & Rhodes, M. (2020). Asking children to "be helpers" can backfire after setbacks. *Child Development*, *91*(1), 236–248. https://doi.org/10.1111/cdev.13147

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, *14*(4), 247–256. https://doi.org/10.1016/0162-3095(93)90020-I

Freeman, R. E., Wicks, A. C., & Parmar, B. (2004). Stakeholder theory and "the corporate objective revisited." *Organization Science*, *15*(3), 364–369. https://doi.org/10.1287/orsc.1040.0066

Futamura, I. (2018). Is extraordinary prosocial behavior more valuable than ordinary prosocial behavior? *PLOS ONE*, *13*(4), e0196340. https://doi.org/10.1371/journal.pone.0196340

Gelfand, M. J., Harrington, J. R., & Jackson, J. C. (2017). The Strength of Social Norms Across Human Groups. *Perspectives on Psychological Science*, *12*(5), 800–809. https://doi.org/10.1177/1745691617708631

Georgeac, O. A. M., Rattan, A., & Effron, D. A. (2019). An exploratory investigation of Americans' expression of gender bias before and after the 2016 presidential election. *Social Psychological and Personality Science*, *10*(5), 632–642. https://doi.org/10.1177/1948550618776624

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107–119. https://doi.org/10.1037/0003-066X.46.2.107

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Harvard University Press.

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

Gneezy, A., & Epley, N. (2014). Worth keeping but not exceeding: Asymmetric consequences of breaking versus exceeding promises. *Social Psychological and Personality Science*, *5*(7), 796–804. https://doi.org/10.1177/1948550614533134

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/S15324834BASP2204_3

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

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

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Gray, K., & Wegner, D. M. (2011). To escape blame, don't be a hero—Be a victim. *Journal of Experimental Social Psychology*, *47*(2), 516–519. https://doi.org/10.1016/j.jesp.2010.12.012

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

Greene, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In *Advances in Experimental Social Psychology* (Vol. 24, pp. 319–359). Elsevier. https://doi.org/10.1016/S0065-2601(08)60333-0

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

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

Haidt, Jonathan. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, Jonathan, & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, *20*(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z

Hamilton, V. L., Blumenfeld, P. C., & Kushler, R. H. (1988). A question of standards: Attributions of blame and credit for classroom acts. *Journal of Personality and Social Psychology*, *54*(1), 34–48. https://doi.org/10.1037/0022-3514.54.1.34

Haney, C., Sontag, L., & Costanzo, 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

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences*, *18*(4), 167–170. https://doi.org/10.1016/j.tics.2014.01.005

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). Guilford Press.

Henderlong, J., & Lepper, M. R. (2002). The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychological Bulletin*, *128*(5), 774–795. https://doi.org/10.1037/0033-2909.128.5.774

Henrich, J. (2009). The evolution of costly displays, cooperation and religion. *Evolution and Human Behavior*, *30*(4), 244–260. https://doi.org/10.1016/j.evolhumbehav.2009.03.005

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission.

Evolution and Human Behavior, *22*(3), 165–196. https://doi.org/10.1016/S1090-5138(00)00071-4

Heyman, G., Barner, D., Heumann, J., & Schenck, L. (2014). Children's sensitivity to ulterior motives when evaluating prosocial behavior. *Cognitive Science*, *38*(4), 683–700. https://doi.org/10.1111/cogs.12089

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

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520–549. https://doi.org/10.1037/xge0000569

Hok, H., Martin, A., Trail, Z., & Shaw, A. (2020). When children treat condemnation as a signal: The costs and benefits of condemnation. *Child Development*, *91*(5), 1439–1455. https://doi.org/10.1111/cdev.13323

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

James, R. N., & Jones, K. S. (2011). Tithing and religious charitable giving in America. *Applied Economics*, *43*(19), 2441–2450. https://doi.org/10.1080/00036840903213384

Janoff-Bulman, R., & Carnes, N. C. (2013). Surveying the moral landscape: Moral motives and group-based moralities. *Personality and Social Psychology Review*, *17*(3), 219–236. https://doi.org/10.1177/1088868313480274

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

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

Kamtekar, R., & Nichols, S. (2019). Agent-regret and accidental agency. *Midwest Studies In Philosophy*, *43*(1), 181–202. https://doi.org/10.1111/misp.12112

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580. https://doi.org/10.1073/pnas.0705345104

Kinzler, Katherine D., & Shutts, K. (2008). Memory for "mean" over "nice": The influence of threat on children's face memory. *Cognition*, *107*(2), 775–783. https://doi.org/10.1016/j.cognition.2007.09.005

Klein, N., & Epley, N. (2014). The topography of generosity: Asymmetric evaluations of prosocial actions. *Journal of Experimental Psychology: General*, *143*(6), 2366–2379. https://doi.org/10.1037/xge0000025

Knafo, A., Schwartz, S. H., & Levine, R. V. (2009). Helping strangers is lower in embedded cultures. *Journal of Cross-Cultural Psychology*, *40*(5), 875–879. https://doi.org/10.1177/0022022109339211

Kohlberg, L. (1981). *The philosophy of moral development*. Harper & Row.

Konrath, S., Meier, B. P., & Bushman, B. J. (2014). Development and validation of the Single Item Narcissism Scale (SINS). *PLoS ONE*, *9*(8), e103469. https://doi.org/10.1371/journal.pone.0103469

Kraft-Todd, G. T., & Rand, D. G. (2019). Rare and costly prosocial behaviors are perceived as heroic. *Frontiers in Psychology*, *10*, 234. https://doi.org/10.3389/fpsyg.2019.00234

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

Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory*, *15*(2), 127–147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x

Lemoine, G. J., Hartnell, C. A., & Leroy, H. (2019). Taking stock of moral approaches to leadership: An integrative review of ethical, authentic, and servant leadership. *Academy of Management Annals*, *13*(1), 148–187. https://doi.org/10.5465/annals.2016.0121

Levine, R. V., Norenzayan, A., & Philbrick, K. (2001). Cross-cultural differences in helping strangers. *Journal of Cross-Cultural Psychology*, *32*(5), 543–560. https://doi.org/10.1177/0022022101032005002

Lindsey, L. L. M. (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

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

Martin, J., Rychlowska, M., Wood, A., & Niedenthal, P. (2017). Smiles as multipurpose social signals. *Trends in Cognitive Sciences*, *21*(11), 864–877. https://doi.org/10.1016/j.tics.2017.08.007

Masicampo, E. J., Barth, M., & Ambady, N. (2014). Group-based discrimination in judgments of moral purity-related behaviors: Experimental and archival evidence. *Journal of Experimental Psychology: General*, *143*(6), 2135–2152. https://doi.org/10.1037/a0037831

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., 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

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

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science*, *31*(3), 227–242. https://doi.org/10.1177/0956797619900321

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

Miller, J. G., & Bersoff, D. M. (1992). Culture and moral judgment: How are conflicts between justice and interpersonal responsibilities resolved? *Journal of Personality and Social Psychology*, *62*(4), 541–554. https://doi.org/10.1037/0022-3514.62.4.541

Miller, J. G., Bersoff, D. M., & Harwood, R. L. (1990). Perceptions of social responsibilities in India and in the United States: Moral imperatives or personal decisions? *Journal of Personality and Social Psychology*, *58*(1), 33–47. https://doi.org/10.1037/0022-3514.58.1.33

Morewedge, C. K., Giblin, C. E., & Norton, M. I. (2014). The (perceived) meaning of spontaneous thoughts. *Journal of Experimental Psychology: General*, *143*(4), 1742–1754. https://doi.org/10.1037/a0036775

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

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology*, *75*(1), 33–52. https://doi.org/10.1037/0022-3514.75.1.33

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

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

Ngo, L., Kelly, M., Coutlee, C. G., Carter, R. M., Sinnott-Armstrong, W., & Huettel, S. A. (2015). Two distinct moral mechanisms for ascribing and denying intentionality. *Scientific Reports*, *5*(1), 17390. https://doi.org/10.1038/srep17390

Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H.-Y. (2016). Rational Learners and Moral Rules: Rational Learners and Moral Rules. *Mind & Language*, *31*(5), 530–554. https://doi.org/10.1111/mila.12119

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 conterfactual thinking. *Journal of Personality and Social Psychology*, *67*(4), 585–595. https://doi.org/10.1037/0022-3514.67.4.585

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

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

Padilla, A., Hogan, R., & Kaiser, R. B. (2007). The toxic triangle: Destructive leaders, susceptible followers, and conducive environments. *The Leadership Quarterly*, *18*(3), 176–194. https://doi.org/10.1016/j.leaqua.2007.03.001

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

Paluck, E. L., & Shepherd, H. (2012). The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network. *Journal of Personality and Social Psychology*, *103*(6), 899–915. https://doi.org/10.1037/a0030015

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*(6), 905. 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

Piaget, J. (1965). *The moral judgment of the child* (M. Gabian, Trans.; Original work published 1932). Free Press.

Pizarro, D. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour*, *30*(4), 355–375. https://doi.org/10.1111/1468-5914.00135

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.), *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

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

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781–799. https://doi.org/10.1037/0033-295X.111.3.781

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75. https://doi.org/10.1037/a0021867

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

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

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104–112. https://doi.org/10.1037/0022-3514.64.1.104

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), 233–240. https://doi.org/10.1037/a0022598

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544. https://doi.org/10.1016/j.cognition.2020.104544

Schein, C., Jackson, J. C., Frasca, T., & Gray, K. (2020). Praise-many, blame-fewer: A common (and successful) strategy for attributing responsibility in groups. *Journal of Experimental Psychology: General*, *149*(5), 855–869. https://doi.org/10.1037/xge0000683

Schmader, T., & Lickel, B. (2006). The approach and avoidance function of guilt and shame emotions: Comparing reactions to self-cause and other-cause 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*. Springer New York. https://doi.org/10.1007/978-1-4612-5094-4

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), 177. https://doi.org/10.2307/1130649

Shweder, R. A., Mahapatra, M., & Miller, J. G. (1987). Culture and moral development. In *The emergence of morality in young children.* (pp. 1–83). University of Chicago Press.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211. https://doi.org/10.1016/j.cognition.2017.05.004

Sims, R. R., & Brinkman, J. (2002). Leaders as moral role models: The case of John Gutfreund at Salomon Brothers. *Journal of Business Ethics*, *35*(4), 327–339. https://doi.org/10.1023/A:1013826126058

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In *Psychology of Learning and Motivation* (Vol. 50, pp. 1–26). Elsevier. https://doi.org/10.1016/S0079-7421(08)00401-5

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development*, *52*(4), 1333. https://doi.org/10.2307/1129527

Smith, H. M. (1991). Varieties of moral worth and moral credit. *Ethics*, *101*(2), 279–303. https://doi.org/10.1086/293289

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<191::AID-JCLP2270280220>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. H. (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*(1), 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Tankard, M. E., & Paluck, E. L. (2017). The effect of a Supreme Court decision regarding gay marriage on social norms and personal attitudes. *Psychological Science*, *28*(9), 1334–1344. https://doi.org/10.1177/0956797617709594

Tapp, J. L., & Kohlberg, L. (1971). Developing senses of law and legal justice. *Journal of Social Issues*, *27*(2), 65–91. https://doi.org/10.1111/j.1540-4560.1971.tb00654.x

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*(7), 1404–1417. https://doi.org/10.1177/0956797614532474

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. *The 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

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476–477. https://doi.org/10.1111/j.1467-9280.2006.01731.x

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*. Guilford Press.

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.

Wiltermuth, S. S., Monin, B., & Chow, R. M. (2010). The orthogonality of praise and condemnation in moral judgment. *Social Psychological and Personality Science*, *1*(4), 302–310. https://doi.org/10.1177/1948550610363162

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), 1–37. https://doi.org/10.1080/10463280600574815

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

Yudkin, D. A., Prosser, A. M. B., & Crockett, M. J. (2019). Actions speak louder than outcomes in judgments of prosocial behavior. *Emotion*, *19*(7), 1138–1147. https://doi.org/10.1037/emo0000514

Yuki, M., Maddux, W. W., Brewer, M. B., & Takemura, K. (2005). Cross-cultural differences in relationship- and group-based trust. *Personality and Social Psychology Bulletin*, *31*(1), 48–62. https://doi.org/10.1177/0146167204271305

Zhao, L., Chen, L., Sun, W., Compton, B. J., Lee, K., & Heyman, G. D. (2020). Young children are more likely to cheat after overhearing that a classmate is smart. *Developmental Science*, *23*(5). https://doi.org/10.1111/desc.12930

Zhao, L., Heyman, G. D., Chen, L., & Lee, K. (2017). Praising young children for being smart promotes cheating. *Psychological Science*, *28*(12), 1868–1870. https://doi.org/10.1177/0956797617721529