



The Consequences of Inaccurate Group Meta-Perception

Citation

Lees, Jeffrey Martin. 2020. The Consequences of Inaccurate Group Meta-Perception. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37368848>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available. Please share how this access benefits you. [Submit a story](#)

The Consequences of Inaccurate Group Meta-Perception

A dissertation presented

by

Jeffrey Martin Lees

to

The Committee for the Ph.D. in Business Studies

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Organizational Behavior

Harvard University

Cambridge, Massachusetts

July, 2020

© 2020 Jeffrey Martin Lees
All rights reserved.

The Consequences of Inaccurate Group Meta-Perception

Abstract

How we believe others perceive us – meta-perception – plays a critical role in how we make reputationally impactful decisions. For social group members, meta-perceptive concerns arise not only at the individual level but at the level of the groups of which they are members. Judgments of how others perceive one’s group – group meta-perception – are an essential aspect of group members’ considerations surrounding group-level behaviors. Nonetheless, the possibility of inaccuracy and bias in group meta-perception represents a challenge to group leaders and members hoping to foster and maintain cooperative group relationships and a positive group reputation. Utilizing experimental, survey, and field data, I explore the (in)accuracy of group meta-perception and its consequences for group members and group-level outcomes. In Chapter 1 I find a robust negativity bias in group meta-perception amongst Democrats and Republicans. I demonstrate that inaccuracy is associated with negative out-group motive attributions and that a simple informational intervention reduces these negative motive attributions. In Chapter 2 I develop a theory of organizational moral prospection by synthesizing scholarship across moral psychology and corporate social (ir)responsibility to understanding how leaders in for-profit organizations fail to predict moral backlash toward organizational decisions. In Chapter 3 I utilize field surveys to assess (in)accuracy and bias in group meta-perception amongst for-profit organizational managers facing decisions in the domain of corporate social responsibility.

Table of Contents

Title Page.....	i
Copyright.....	ii
Abstract.....	iii
Table of Contents.....	iv
Acknowledgements.....	v
Introduction.....	1
Chapter 1: Inaccurate Group Meta-Perceptions Drive Negative Out-Group Attributions in Competitive Contexts.....	6
Chapter 2: Moral Prospection: Cognitive Bias and The Failure to Predict Moral Backlash Towards an Organization.....	34
Chapter 3: Do Managers Know How Their Decisions Are Perceived? Measuring the Accuracy of CSR-Related Group Meta-Perception.....	63
Supplementary Notes.....	80
Appendix.....	93
References.....	103

Acknowledgments

I am incredibly grateful to my mentors on my dissertation committee: Francesca Gino, Mina Cikara, and Adam Waytz. Francesca Gino taught me how to manage a research pipeline and always consider how my research relates to practice. Without her constant guidance and willingness to speak on any issue I was facing I would not have finished this dissertation. Mina Cikara taught me how to translate a research idea into a published paper. She also helped me see how a phenomenon can be understood through multiple disciplinary lenses and in multiple contexts, something that is reflected in the work presented herein. Adam Waytz taught me how to connect abstract cognitive processes to real world phenomena. Through him I also learned how to successfully manage long-distance collaborations, a skill I hope to carry forward throughout my career. Most of all, I am grateful to my mentors for their willingness to support me and my development throughout my doctoral career. Whether taking me on as an advisee, inviting me to be part of their lab, or a willingness to collaborate after nothing more than a cold email, I would not be the scholar I am today without their faith in me throughout my growth as a scholar. I am also deeply indebted to my many collaborators on projects outside the scope of this dissertation, including Liane Young, Evan Apfelbaum, Jim Sidanius, Simone Tang, and Katy DeCelles.

I am exceedingly grateful to my family for their support throughout my doctoral studies. I cannot thank my wife Rebecca Young enough for her loving, emotional support over these five years, and help in my efforts to become a better writer. Without her I could not have been successful in any of my scholarly endeavors. I am also thankful to my parents, Rick and Kathy, and siblings, Kevin and Andrea, for their support and enthusiasm as I pursued this career. They

have served as an anchor, helping to remind me that there is more to life than purely academic pursuits.

I owe a huge debt to my peers at Harvard. My fellow students in OB Lab, the Harvard Intergroup Neuroscience Lab, the Sidanius Lab, and my doctoral cohort at HBS have been tremendously supportive, socially and academically. Without their constructive feedback and social support over the past five years I cannot imagine how I would have been able to thrive.

I am thankful for the generous financial support over the past five years from the Harvard Business School Doctoral Office, Harvard Graduate School of Arts and Sciences, and The Foundations of Human Behavior Initiative. All the data collected as part of this dissertation were supported by these funding sources, and could not have been completed without their generosity.

Introduction

Meta-perceptive accuracy has long been of interest to psychology (Carlson et al., 2011; Kenny & DePaulo, 1993; Laing et al., 1966). Yet because it has been embedded in the literature on interpersonal perception, meta-perception and questions of meta-accuracy have been studied almost exclusively at the individual-level, such as meta-perception in the domain of personality judgments (Vazire, 2010), close relationships (Carlson, 2016), the workplace (Eisenkraft et al., 2017), stereotyping (Vorauer et al., 2000), and prejudice (Kteily et al., 2016). Yet concerns regarding the perceptions of others are not limited to perceptions toward the self, they extend to perceptions toward the collective behaviors of social groups.

Social judgments of groups, as discrete agents in and of themselves, can differ both from judgments of individual agents and judgments made of group members (Ames, 2004; Jago & Laurin, 2017; Lickel et al., 2001; Rai & Diermeier, 2015; Waytz & Young, 2012). And while reputation concerns, both within and between groups, have long been a focus of social science research (Emler, 1990; Lange et al., 2011; Semmann et al., 2005; Solomon & Vazire, 2016), whether group members can accurately assess their own group's reputation, and how specific group-level behaviors may affect the group's reputation among those outside the group, has received scant empirical and theoretical investigation. The possibility of inaccuracy and systematic biases in such group meta-perceptive judgments presents a host of concerns, across intergroup and organizational contexts, related to group-level decision-making and the maintenance of cooperative group-level relationships.

I approach group meta-perception as a cognitive process which, like individual meta-perception, I conceived of as a domain-general process individuals engage in across social contexts, and which represents a specific instantiation of broader processes such as prospection

and theory of mind (Knobe, 2005; Seligman et al., 2013). Across Chapters 1-3 I argue that group meta-perception can be systematically inaccurate, and that inaccuracy drives negative outcomes for groups and organizations attempting to maintain cooperative relationships and positive reputations.

Overview of Chapter 1

Chapter One, published in *Nature Human Behaviour* (Lees & Cikara, 2020), provides empirical evidence, across six studies, that inaccurate group meta-perceptions drive negative out-group attributions in the domain of intergroup relations, while showing that a simple intervention which corrects these inaccurate beliefs can reduce such negative attributions. In Study 1 I demonstrate that both Democrats and Republicans have highly inaccurate and overly negative beliefs about how the out-group will perceive their party's behavior. Study 1 also finds that this negativity bias persists even when participants are forecasting the reactions of hypothetical and anonymized political parties, suggesting that the negativity bias represents how we simulate group-on-group competition rather than just how individuals perceive their local out-group. In Study 2 I demonstrate that the negativity bias in group meta-perception generalizes to the context of gender-relations. Study 3 returns to the context of US politics and finds that when the interaction between Democrats and Republicans is cooperative rather than competitive group meta-perceptions become accurate. This finding suggests that the observed inaccuracies in Studies 1 and 2 represent a negativity bias related to group-competition, rather than an extremity bias in how we forecast out-group reactions across contexts where an out-group may have a plausibly positive or negative reaction to the behavior of the in-group.

Study 4 replicates the negativity bias observed in Study 1 with a nationally representative sample and using a repeated-measures design. Moreover, Study 4 finds that individuals'

forecasts of how other in-group members will perceive competitive intergroup behavior is more negative than how those individuals themselves felt, suggesting that the negativity bias in group meta-perception is not exclusive to how we cognitively simulate the out-group. Study 5 finds that those who have more inaccurate and negative group meta-perceptions are also more likely to make negative motive attributions toward the out-group. In an effort to reduce such negative motive attributions, Study 6 employs an experimental, informational intervention which significantly reduces such attributions and is more effective on those who exhibit greater baseline inaccuracy. Chapter 1 not only provides insights into ways of reducing negative intergroup attitudes, but also highlights the importance of pernicious biases in group meta-perceptive judgments and demonstrates how such biases can lead to negative outcomes.

Overview of Chapter 2

In Chapter 2 I introduce a theoretical model which attempts to elucidate the cognitive mechanisms underlying why organizational leaders fail to anticipate moral backlash toward organizational decisions. I first define the process of moral prospection: the act of attempting to estimate and simulate how organizational outsiders will perceive a potential decision, on the part of the organization, in the moral domain. Drawing from research on anthropomorphization I argue the process of moral prospection among organizational insiders is subject to a corporate personhood bias, where insiders who fully anthropomorphize their organization fail to anticipate reduced levels of anthropomorphization among outsiders when simulating the moral attributions outsiders will make of a given organizational decision. This bias leads insiders to generate inaccurate and overly positive group moral meta-perceptions in specific attributional domains, leading to a state I call moral overconfidence. Moral overconfidence contributes to a process whereby insiders update and codify their own moral preferences surrounding the decision at

hand, ultimately leading to a decision which can engender unexpected moral backlash from relevant stakeholders. Chapter 2 provides a broad theoretical framework for understanding how inaccurate group meta-perceptions can drive organizational scandals, and highlights a set of novel psychological and organizational phenomena.

Overview of Chapter 3

In Chapter 3 I draw from a sample of real-world organizational managers and directly measure their group meta-perceptions in the domain of corporate social responsibility. I first elicit from managers written accounts of decisions they have faced related to corporate social responsibility, along with asking for their group meta-perceptions of how the public would perceive their organization in relation to their decision. Additionally, I measure personality traits to test whether these moderate accuracy in group meta-perception. I then compare these organizational leaders' meta-perceptions to actual-perceptions collected from a general population sample, allowing for a direct test of accuracy and bias in group meta-perceptions, along with moderators of accuracy and outcomes predicted by accuracy.

Summary

Group meta-perception is a vital aspect of group-on-group interaction, and as such inaccuracy and bias in group meta-perception can negatively effect the tenor of group relations and lead to decisions which harm groups' reputations and engender conflict. Chapters 1-3 provide evidence for group meta-perceptive inaccuracy and bias, across intergroup and business contexts, and highlight the consequences of such misperceptions. Chapters 1-3 also help us better understand the cognitive mechanisms underlying group meta-perception, and ways in which accuracy can be increased. Nonetheless, there is much work needed to better understand precisely why, when, and how group meta-perception are (in)accurate. As such, in the Future

Directions section I discuss gaps in our theoretical and empirical knowledge of group meta-perception and how scholars may go about building toward a cumulative science of group meta-perception and its consequences.

Chapter 1

Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts

Jeffrey Lees

Mina Cikara

Published in *Nature Human Behaviour*

DOI: <https://doi.org/10.1038/s41562-019-0766-4>

Abstract

Across seven experiments and one survey (N=4282) people consistently overestimated out-group negativity towards the collective behavior of their in-group. This negativity bias in group meta-perception was present across multiple competitive (but not cooperative) intergroup contexts, and appears to be yoked to group psychology more generally; we observed negativity bias for estimation of out-group, anonymized-group, and even fellow in-group members' perceptions. Importantly, in the context of American politics greater inaccuracy was associated with increased belief that the out-group is motivated by purposeful obstructionism. However, an intervention that informed participants of the inaccuracy of their beliefs reduced negative out-group attributions, and was more effective for those whose group meta-perceptions were more inaccurate. In sum, we highlight a pernicious bias in social judgments of how we believe 'they' see 'our' behavior, demonstrate how such inaccurate beliefs can exacerbate intergroup conflict, and provide an avenue for reducing the negative effects of inaccuracy.

Main

How we believe others perceive us—meta-perception—plays a critical role in how we interact with others (Carlson, 2016; Carlson et al., 2011; Vazire & Carlson, 2011). In the context of intergroup interactions, these meta-perceptions may bring unpleasant, even harmful evaluations to mind (Frey & Tropp, 2006; Kteily et al., 2016; Sigelman & Tuch, 1997; Vorauer et al., 1998, 2000). For example, when individuals believe they are being negatively stereotyped by an out-group member, they experience increased negative emotions and lower self-esteem (Vorauer et al., 1998), suffer increased anxiety (Finchilescu, 2010), and subsequently exhibit more intergroup bias (Klein & Azzi, 2001).

Despite the important role that beliefs about how ‘they’ see ‘us’ (and our actions) (Goldstein et al., 2014; Lau et al., 2016; Saguy & Kteily, 2011; Waytz et al., 2014), past work has focused primarily on person-to-person interactions across group boundaries or on estimates of extremity of and polarization in out-group attitudes (Chambers et al., 2006; Chambers & Melnyk, 2006; Robinson et al., 1995; Westfall et al., 2015). As an example of the latter, findings in the domain of values and attitudes indicate that group members overestimate the level of disagreement and polarization between groups (though note that these constitute first order judgments, or “how I see X”) (Chambers et al., 2006; Chambers & Melnyk, 2006; Westfall et al., 2015). Evidence from the intergroup literature, more broadly, suggests that group labels exacerbate inaccuracy in social judgments because they activate stereotypes that cause people to adjust their judgments away from their initial, more accurate anchors (Lau et al., 2016).

There is, in complement, a growing literature in the domain of second-order, intergroup meta-judgments (or “what I think they think about us”), which reveals that people tend to have overly negative and inaccurate judgments of out-group motives toward the in-group (Saguy &

Kteily, 2011; Waytz et al., 2014). This foundational work on the effects of meta-perceptions in intergroup contexts raises two important questions: (i) are these meta-perceptions accurate?; and (ii) what happens when these judgments are made in response to collective action—when people consider how ‘they’ see ‘our’ (not my) behavior?

Here, we tackle a particular form of intergroup inaccuracy by examining group meta-perceptions (GMPs): how we believe our group’s collective actions will be perceived by the out-group. In our view, GMPs represent an intergroup-context activated distortion of second-order judgments. This makes GMPs (i) distinct from first-order judgments and (ii) unique in that they should be sensitive to functional relations between groups (i.e., whether groups are cooperative, competitive, etc.) but relatively invariant to the focal event/act/behavior or the groups in question.

GMPs likely serve an important role in determining the course of group-on-group interaction because they allow us to make predictions about whether an out-group will be supportive or hostile towards our own group’s efforts at cooperation; therefore, GMPs should also drive emotions, strategy, and policy preferences. For example, U.S. President George W. Bush, in his address to a joint session of Congress on September 20th, 2001 laid out in stark terms how he believed Al-Qaeda perceived the United States, and how these second-order judgments ought to compel US foreign policy (Bush, 2001): “Americans are asking ‘Why do they hate us?’ They hate what they see right here in this chamber: a democratically elected government...They hate our freedoms: our freedom of religion, our freedom of speech, our freedom to vote and assemble and disagree with each other...We will direct every resource at our command—every means of diplomacy, every tool of intelligence, every instrument of law enforcement, every financial influence, and every necessary weapon of war—to the destruction

and to the defeat of the global terror network....Every nation in every region now has a decision to make: Either you are with us or you are with the terrorists.” President Bush used the belief that “they hate our freedoms” to motivate his call to war and his ultimatum to other countries that they are either “with us” or “with the terrorists.” However, many have noted that this belief that Al-Qaeda “hate our freedoms” wrongly diagnosed the motivations of Al-Qaeda and the complex socio-political forces which drove their perception of the United States (Sunstein, 2002; Zakaria, 2001). Furthermore, this essentializing language served to dehumanize Muslims and drive support for the “War on Terror” among the American public (Merskin, 2004).

This example highlights how inaccurate, and overly negative, beliefs about how the out-group perceives the behavior (and values) of one’s own group can drive intractable intergroup conflict. When group leaders and other group members believe the out-group will react with animosity and perceive one’s group in a highly negative fashion, they are likely to support antagonistic intergroup actions over cooperative and reconciliatory behaviors. For example, when people believe they are dehumanized by an out-group, they are more likely to dehumanize the out-group in return, which leads to increased support for war and out-group torture (Kteily et al., 2016). This dynamic can unfold in contexts as hostile as war between nations, but also legislative compromise across political parties, competitive sports, and interaction between organizations. Nonetheless, interventions which directly inform individuals of their inaccurate beliefs may be able to induce positive behavioral change (Nyhan & Reifler, 2018; Rogers & Feller, 2018).

To investigate the nature of GMPs we constructed a set of scenarios involving group-level conflict. For Experiments 1, 3, 4, 6 and Study 5 these scenarios pertained to the behavior of American political parties in a legislative context. In Experiment 2 the scenarios pertained to

group-level conflict between men and women in educational and workplace settings. All scenarios presented instances where one group was attempting to pass a law or change a policy in a manner which would potentially disadvantage the other group (e.g., requiring a sitting governor of the opposing party to disclose their taxes), except for Experiment 3 where the behavior would potentially benefit the other group. Supplemental Experiment A is a direct replication of Experiment 4 with a convenience sample, and Supplemental Experiment B is an exploratory follow up to Experiment 6.

Experiments 1-4 were designed to test for participant accuracy in GMPs. At the beginning of these experiments participants were asked to identify their political affiliation (or gender identity in Experiment 2) and were then randomly assigned to whether the group taking action in the scenario was their in-group or out-group. Those who read about their in-group taking action were asked for their GMPs (e.g., “How much do you believe an [out-group member] will dislike this action?”), whereas those who read about their out-group taking action against their in-group were asked for their actual perceptions (e.g., “How much do you dislike this action?”). In Experiment 4 we also asked about “in-group perceptions” (e.g., “How much do you believe an [in-group] member will dislike the [out-group] action?”). Across all experiments, the comparison of the GMP and actual perception conditions across groups (that is, Democrats vs. Republicans and men vs. women) allowed for a direct test of participant accuracy.

When reading the scenarios participants were asked, either as a meta-perception, actual perception, or in-group perception, their perceived dislike of, opposition to, and political/social unacceptability of the action being taken in the scenario, which they reported on sliding scales, with labels at the end of the scales (e.g., 1=“Not Opposed”, 100=“Extremely Opposed”). After the ratings all participants, across all experiments and Study 5, completed a comprehension

check which asked them to identify the group “taking action” in the scenario. Any participants who failed this check were excluded from all analyses. Lastly, all participants were asked their age, gender, and whether they had comments for the experimenters (except in Experiment 4 in which demographic questions were asked at the beginning of the experiment).

All materials, data, and analysis code for all experiments and studies, and preregistrations for Experiments 4 and 6, are available on OSF: <https://osf.io/zhysa/>

Results

Experiments 1-4 and Study 5 were analyzed using mixed-effects beta-regressions and Experiment 6 was analyzed using linear mixed-effects regression. All tests are two-sided. In Experiment 6 homoscedasticity and normality of errors was assumed but was not formally tested. Further details regarding the analyses can be found in the Methods section.

In Experiment 1 (N=408), participants were randomly assigned to the GMP condition (N=129), actual-perception condition (N=143), or an unlabeled and anonymized control group meta-perception condition (N=136) where participants were asked how “Party B” would perceive the behavior of “Party A.” Within each condition, participants were randomly assigned to read one of five scenarios (we included multiple scenarios in each experiment and study to assess the robustness of our effects and modeled scenario as a random effect).

Across all scenarios, participants in the GMP condition substantially overestimated the negative perceptions of out-group participants (i.e., out-group members in the actual-perception condition) on our three measures: action dislike (unstandardized log-odds regression coefficient (b)=1.51, 95% confidence interval (CI)=[1.19,1.83], odd-ratio (OR)=4.53, Z-score (z)=9.27, $P < 0.001$), opposition to the action (b =1.40, 95% CI=[1.09,1.72], OR=4.08, z =8.78, $P < 0.001$), and political unacceptability of the action (b =1.36, 95% CI=[1.04,1.67], OR=3.89, z =8.46, $P <$

0.001). Similarly, participants in the control meta-perception condition overestimated the negative perceptions of those in the actual-perception condition: dislike ($b=1.32$, 95% CI=[1.02,1.62], OR=3.74, $z=8.55$, $P < 0.001$), opposition ($b=1.22$, 95% CI=[0.93,1.52], OR=3.40, $z=8.15$, $P < 0.001$), and political unacceptability ($b=1.13$, 95% CI=[0.83,1.42], OR=3.08, $z=7.45$, $P < 0.001$). Pairwise post-hoc tests indicate no statistically significant difference between responses in the control meta-perception condition vs. the GMP condition: dislike ($b=-0.19$, 95% CI=[-0.54,0.15], OR=0.83, $t(402)=-1.30$, $P=0.40$), opposition ($b=-0.18$, 95% CI=[-0.52,0.16], OR=0.83, $t(402)=-1.24$, $P=0.43$), and political unacceptability ($b=-0.23$, 95% CI=[-0.58,0.11], OR=0.79, $t(401)=-1.58$, $P=0.26$). We also examined the main effect of accuracy by party, modeled as a categorical fixed effect with two groups: “Democrat Accuracy”—Democrats in the GMP and control conditions compared with Republicans in the actual-perception condition—and “Republican Accuracy”—Republicans in the GMP and control conditions compared with Democrats in the actual-perception condition (see Methods for model details). This approach allowed the main-effect to appropriately contrast meta/control vs. actual perceptions (the baseline in the analyses) across parties, rather than within party. Indeed, there was no statistically significant main effect of party accuracy: dislike ($b=-0.04$, 95% CI=[-0.29,0.21], OR=0.96, $z=-0.32$, $P=0.75$), opposition ($b=-0.00$, 95% CI=[-0.25,0.24], OR=1.00, $z=-0.03$, $P=0.98$), and political unacceptability ($b=-0.02$, 95% CI=[-0.27,0.23], OR=0.98, $z=-0.18$, $P=0.85$). Finally, pairwise post-hoc tests found no statistically significant differences when examining whether Democrats and Republicans differed in their actual perceptions of the scenarios: dislike ($b=0.00$, 95% CI=[-0.61,0.61], OR=1.00, $t(400)=0.02$, $P=1.00$), opposition ($b=0.15$, 95% CI=[-0.45,0.74], OR=1.16, $t(400)=0.72$, $P=0.98$), and political unacceptability

($b=0.11$, 95% CI=[-0.49,0.71], OR=1.11, $t(399)=0.51$, $P=1.00$). See Figure 1.1 for a visualization of the raw data by condition.

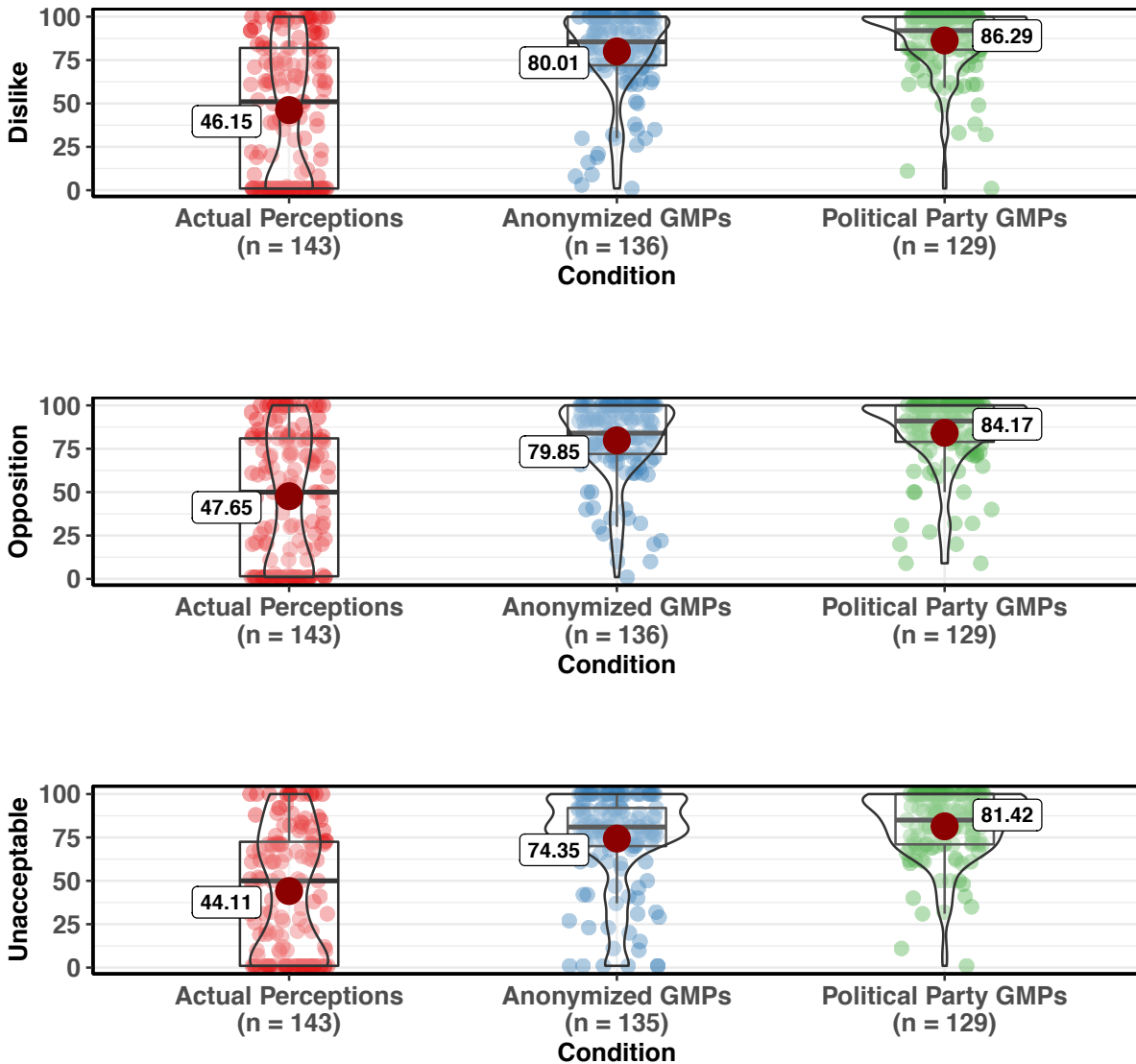


Figure 1.1: Raw data from Experiment 1 by condition and dependent variable. In this experiment, $N=408$ (collected via Mechanical Turk). In the two GMP conditions participants reported how much they thought their out-group, or an anonymized political party (control), would dislike, oppose, and find unacceptable the in-group's/other party's action in the scenario. Solid red dots and corresponding numbers are sample means, the boxplot center lines are sample medians. Participants in the political party GMP condition overestimated the negative perceptions of out-group participants in the actual-perception condition on action dislike ($b=1.51$, 95% CI=[1.19,1.83], OR=4.53, $z=9.27$, $P < 0.001$), opposition to the action ($b=1.40$, 95% CI=[1.09,1.72], OR=4.08, $z=8.78$, $P < 0.001$), and political unacceptability of the action ($b=1.36$, 95% CI=[1.04,1.67], OR=3.89, $z=8.46$, $P < 0.001$). Participants in the control meta-perception condition overestimated the negative perceptions of those in the actual-perception condition on dislike ($b=1.32$, 95% CI=[1.02,1.62], OR=3.74, $z=8.55$, $P < 0.001$), opposition ($b=1.22$, 95%

CI=[0.93,1.52], OR=3.40, $z=8.15$, $P < 0.001$), and political unacceptability ($b=1.13$, 95% CI=[0.83,1.42], OR=3.08, $z=7.45$, $P < 0.001$). Pairwise post-hoc tests indicate no statistically significant difference between responses in the control meta-perception vs. GMP condition on dislike ($b=-0.19$, 95% CI=[-0.54,0.15], OR=0.83 $t(402)=-1.30$, $P=0.40$), opposition ($b=-0.18$, 95% CI=[-0.52,0.16], OR=0.83, $t(402)=-1.24$, $P=0.43$), and political unacceptability ($b=-0.23$, 95% CI=[-0.58,0.11], OR=0.79, $t(401)=-1.58$, $P=0.26$). These results provide evidence of overly pessimistic GMPs.

As predicted, GMPs in Experiment 1 were more negative than participants' actual perceptions of the out-group's behavior. This was true even when we removed party labels. Thus, merely invoking the political intergroup context was enough to engender inaccuracy, supporting our proposition that GMPs are an intergroup-context activated distortion, invariant to the groups in question. Furthermore, we found no credible evidence that this effect was moderated by participants' party membership. This suggests that Democrats and Republicans were equally pessimistic, and therefore inaccurate, in judging how members of the other party perceived the collective behavior of their own party.

To further examine the generalizability of our findings, Experiment 2 (N=286) utilized a design similar to that of Experiment 1, but in the context of gender relations. There were two changes from the design of Experiment 1. First, participants were assigned to one of three scenarios regarding group-level gender conflict (e.g., integrating a single-gender school choir), rather than five scenarios regarding political conflict. Second, we did not include an anonymized-group control condition. As with Experiment 1, participants were randomly assigned to the GMP condition (N=128) or actual perception condition (N=158), read only one scenario, and responded to items regarding perceived dislike of, opposition to, and social unacceptability of the action in the scenario.

Results indicated a statistically significant condition (actual vs. meta perception) by gender-accuracy interaction (i.e., a fixed effect similar to “party accuracy” in Experiment 1, contrasting accuracy across gender rather than within gender), indicating that one gender had less

inaccurate GMPs than the other: dislike ($b=0.78$, 95% CI=[0.22,1.34], OR=2.18, $z=2.73$, $P=0.006$), opposition ($b=0.74$, 95% CI=[0.18,1.30], OR=2.09, $z=2.59$, $P=0.010$), and social unacceptability ($b=0.65$, 95% CI=[0.09,1.21], OR= 1.92, $z=2.27$, $P=0.023$). Pairwise post-hoc tests revealed that female participants had highly negative and inaccurate GMPs, replicating Experiment 1: dislike ($b=-1.13$, 95% CI=[-1.66,-0.59], OR=0.32, $t(280)=-5.42$, $P < 0.001$), opposition ($b=-1.07$, 95% CI=[-1.60,-0.54], OR=0.34, $t(280)=-5.22$, $P < 0.001$), and social unacceptability ($b=-1.02$, 95% CI=[-1.56,-0.49], OR=0.36, $t(280)=-4.93$, $P < 0.001$). However, male participants' GMPs were not significantly different from the actual perceptions of female participants: dislike ($b=-0.35$, 95% CI=[-0.86,0.17], OR=0.71, $t(280)=-1.74$, $P=0.30$), opposition ($b=-0.33$, 95% CI=[-0.85,0.20], OR=0.72, $t(280)=-1.69$, $P=0.33$), and social unacceptability ($b=-0.37$, 95% CI=[-0.89,0.14], OR=0.69, $t(280)=-1.87$, $P=0.24$). This interaction was driven by gender differences in actual perceptions. Pairwise post-hoc tests indicated that male and female participants' GMPs were not significantly different across dislike ($b=0.29$, 95% CI=[-0.25,0.82], OR=1.33, $t(280)=1.39$, $P=0.51$), opposition ($b=0.19$, 95% CI=[-0.34,0.73], OR= 1.21, $t(280)=0.91$, $P=0.80$), and social unacceptability ($b=0.52$, 95% CI=[-0.02,1.07], OR=1.69, $t(280)=2.49$, $P=0.063$). However, women's (relative to men's) actual perceptions of the behaviors were significantly more negative across disliking ($b=1.07$, 95% CI=[0.56,1.58], OR=2.91, $t(280)=5.39$, $P < 0.001$), opposition ($b=0.93$, 95% CI=[0.42,1.43], OR=2.53, $t(280)=4.75$, $P < 0.001$), and social unacceptability ($b=1.18$, 95% CI=[0.66,1.69], OR=3.24, $t(280)=5.91$, $P < 0.001$).

Thus, while we found no credible evidence that men's group meta-perceptions about how upset women would be were inaccurate, women's GMPs were inaccurate and overly negative, replicating the results from Experiment 1 in the domain of gender. It is important to reiterate,

however, that the men's 'accuracy' result was driven by differences in male and female participant's actual-perceptions. In other words, men's GMPs were closer to women's actual perceptions because women reported being more upset about the policy changes than men did. This pattern is likely the result of real-world power differences between the genders: men may be marginally less impacted and therefore less upset by disadvantageous policies in the contexts featured in our scenarios. More generally, Experiments 1 and 2 demonstrated GMP inaccuracy, but only as it pertained to the out-group in competitive or zero-sum contexts. To examine whether GMPs reflect a negativity bias or a valence-independent extremity bias, Experiment 3 contrasted GMPs versus actual-perceptions in response to cooperative rather than competitive behaviors.

Experiment 3 (N=499) utilized the same design as the GMP and actual-perception conditions from Experiment 1. While the scenarios pertained to the same political content, the nature of the behaviors was inverted such that the groups were taking cooperative actions, which either benefited the other group or disadvantaged the group taking the action. For example, instead of trying to make equal a partisan redistricting board controlled by the other party, in Experiment 3 the party taking action was trying to make equal a partisan redistricting board controlled by their own party. Participants in the GMP (N=233) and actual-perception (N=266) conditions were asked for their positive perceptions (e.g., 1="Not Supportive", 100="Extremely Supportive"), rather than negative perceptions. Otherwise the procedure was the same as Experiment 1, including the between-subjects random assignment to both condition and scenario.

In contrast to Experiments 1 and 2, Experiment 3 found no credible evidence for GMP inaccuracy in cooperative contexts across the support ($b=-0.02$, 95% CI=[-0.25,0.21], OR=0.98, $z=-0.20$, $P=0.84$), liking ($b=0.12$, 95% CI=[-0.11,0.35], OR=1.13, $z=1.02$, $P=0.31$), or political

acceptability ($b=-0.05$, 95% CI=[-0.28,0.18], OR=0.95, $z=-0.42$, $P=0.67$) measures. There was a main effect (but never an interaction) of party-accuracy for support ($b=0.44$, 95% CI=[0.20,0.67], OR=1.55, $z=3.69$, $P < 0.001$), liking ($b=0.48$, 95% CI=[0.25,0.71], OR=1.61, $z=4.05$, $P < 0.001$), and political acceptability ($b=0.52$, 95% CI=[0.29,0.75], OR=1.69, $z=4.46$, $P < 0.001$), such that Democrats' positive reactions were slightly higher than those of Republicans. GMPs for both parties accurately tracked this mean-level difference. The findings from Experiment 3 parallel other work demonstrating that dyadic meta-perceptions are more accurate when two people are cooperative, but less so when competing (Eisenkraft et al., 2017). Broadly, Experiment 3 also provides evidence that GMP inaccuracy represents specifically a negativity bias in competitive contexts, rather than an extremity bias in how we believe the out-group will react to the in-group's actions in general.

Experiments 1, 2, and 3 are limited in several notable ways. First, they all utilize convenience samples (i.e., Mechanical Turk workers), and as such do not represent general population GMPs and actual-perceptions. Second, the previous experiments do not tell us whether people are inaccurate specifically about how the out-group sees the in-group's behavior or, more generally, how any group sees any other group's behavior. Experiment 4, a preregistered (see OSF: <https://osf.io/atck5>) extension of Experiment 1, utilized a nationally-representative sample and included an in-group perception condition to address these limitations.

Experiment 4 (N=536) featured the same scenarios from Experiment 1. Participants were randomly assigned, between-subjects, to the actual perception condition (N=170), GMP condition (N=206), both of which were the same as Experiment 1, or a new condition called the in-group perception condition (N=160). Participants in the in-group perception condition read the

same scenarios as those in the actual perception condition, but instead of being asked for their individual perceptions they were asked how they believed “another [in-group member]” would perceive the scenarios. In contrast to Experiment 1, participants read and responded to all five scenarios (a repeated-measures factor, modeled as a random effect for participant).

Experiment 4 revealed statistically significant differences between all three conditions on all three outcome measures (see Figure 1.2, continued, for raw data distributions). Actual perceptions were lower than in-group perceptions for opposition ($b=-0.26$, 95% CI=[-0.43,-0.09], OR=0.77, $z=-2.93$, $P=0.003$), unacceptability ($b=-0.25$, 95% CI=[-0.43,-0.07], OR=0.78, $z=-2.72$, $P=0.007$), and disliking ($b=-0.34$, 95% CI=[-0.52,-0.17], OR=0.71, $z=-3.93$, $P < 0.001$). GMPs were higher than in-group perceptions for opposition ($b=0.51$, 95% CI=[0.35,0.68], OR=1.67, $z=6.10$, $P < 0.001$), unacceptability ($b=0.43$, 95% CI=[0.25,0.60], OR=1.53, $z=4.87$, $P < 0.001$), and disliking ($b=0.41$, 95% CI=[0.24,0.57], OR=1.50, $z=4.83$, $P < 0.001$). The pairwise post-hoc contrasts between actual-perceptions and GMPs were also significant for opposition ($b=-0.77$, 95% CI=[-0.97,-0.58], OR=0.46, $t(2669)=-9.27$, $P < 0.001$), unacceptability ($b=-0.67$, 95% CI=[-0.87,-0.47], OR=0.51, $t(2669)=-7.83$, $P < 0.001$), and disliking ($b=-0.75$, 95% CI=[-0.95,-0.56], OR=0.47, $t(2669)=-9.04$, $P < 0.001$), directly replicating the main finding of inaccurate GMPs from Experiment 1, but this time in a nationally representative sample. We also

performed a direct replication of Experiment 4 using a convenience sample (again Mechanical Turk workers) and found practically identical results (see “Supplemental Experiment A”).

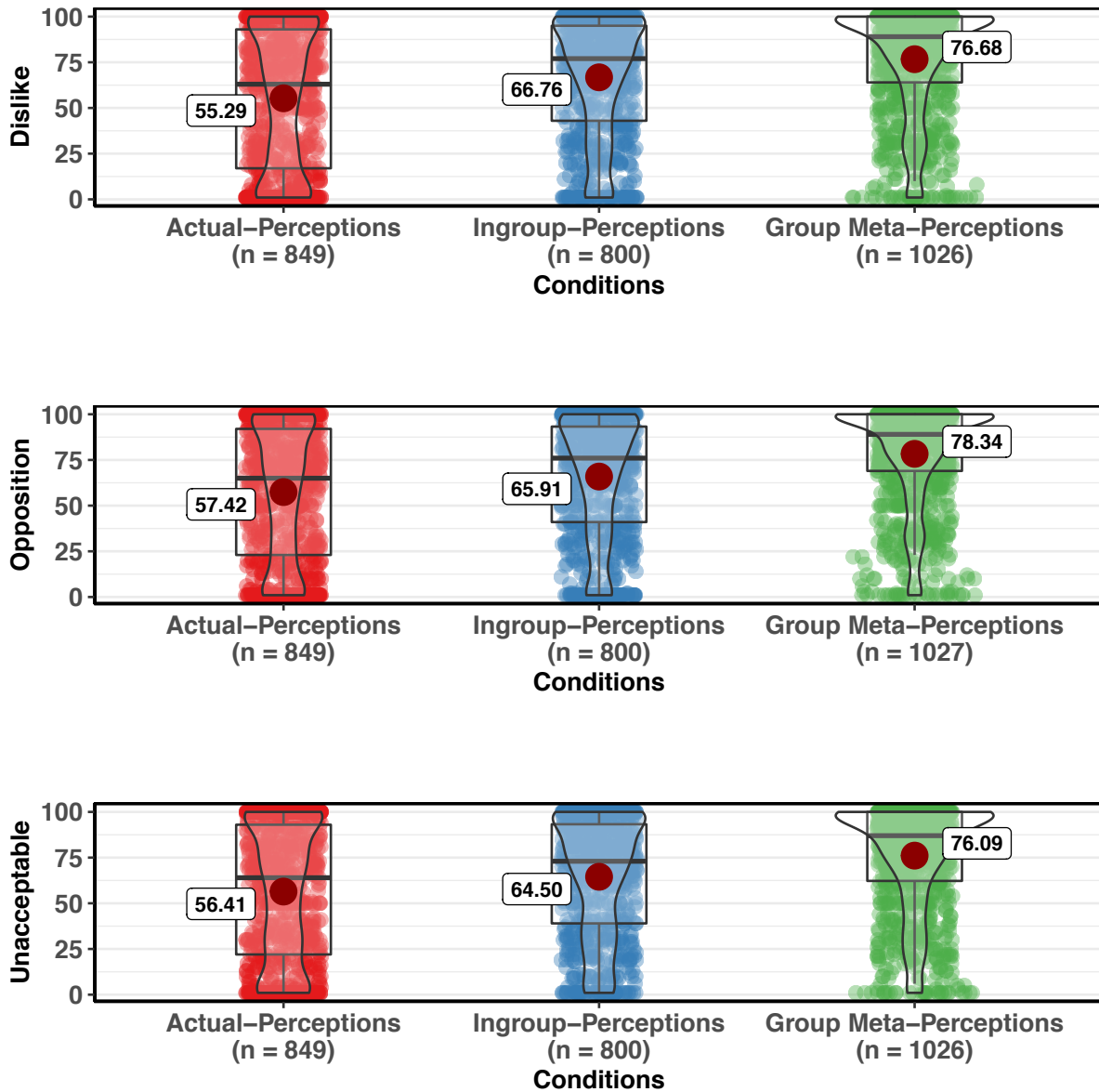


Figure 1.2: Raw data from Experiment 4 by condition and dependent variable. Sample sizes listed in figures are the number of judgments (across five repeated measures). Total N=538 (nationally representative sample collected via Qualtrics survey panels). By Condition: Actual Perceptions N=170, Ingroup Perception N=160, GMPs=206. Solid red dots and corresponding numbers are sample means, the boxplot center lines are sample medians. Actual perceptions were lower than in-group perceptions for opposition ($b=-0.26$, 95% CI=[-0.43,-0.09], OR=0.77, $z=-2.93$, $P=0.003$), unacceptability ($b=-0.25$, 95% CI=[-0.43,-0.07], OR=0.78, $z=-2.72$, $P=0.007$), and disliking ($b=-0.34$, 95% CI=[-0.52,-0.17], OR=0.71, $z=-3.93$, $P < 0.001$). GMPs were higher than in-group perceptions for opposition ($b=0.51$, 95%

CI=[0.35,0.68], OR=1.67, $z=6.10$, $P < 0.001$), unacceptability ($b=0.43$, 95% CI=[0.25,0.60], OR=1.53, $z=4.87$, $P < 0.001$), and disliking ($b=0.41$, 95% CI=[0.24,0.57], OR=1.50, $z=4.83$, $P < 0.001$). The pairwise post-hoc contrasts between actual-perceptions and GMPs were also significant for opposition ($b=-0.77$, 95% CI=[-0.97,-0.58], OR=0.46, $t(2,669)=-9.27$, $P < 0.001$), unacceptability ($b=-0.67$, 95% CI=[-0.87,-0.47], OR=0.51, $t(2,669)=-7.83$, $P < 0.001$), and disliking ($b=-0.75$, 95% CI=[-0.95,-0.56], OR=0.47, $t(2,669)=-9.04$, $P < 0.001$). These results provide evidence of overly pessimistic GMPs and overly pessimistic judgments of the in-group's reactions.

Critically, the differences between in-group perceptions and GMPs indicate that our inaccuracy findings for Experiments 1 and 2 cannot be explained entirely by the difference in referents across the actual perception judgments (“how would you feel”) versus GMP (“how would an out-group member feel”) judgments. In Experiment 4 the in-group judgment also uses a group-level referent (“how would an in-group member feel about the out-group's action”) but is still significantly less negative than the GMP judgments.

Study 5 (N=212) tested whether inaccurate GMPs are consequential by examining the relationship between GMPs and negative motive attributions towards the out-group. In this study, participants completed the GMP condition from Experiment 1. They then reported how much they agreed with the statement “[Out-group members] are purposefully obstructing the process surrounding the [specific scenario topic]” (1-100 slider scale, “Strongly Disagree” to “Strongly Agree”). Analyses indicated a significant positive linear association between the belief that the out-group is obstructionist and negative GMPs of disliking ($b=2.12$, 95% CI=[1.40,2.84], OR=8.34, $z=5.76$, $P < 0.001$), opposition ($b=1.95$, 95% CI=[1.19,2.70], OR=7.00, $z=5.06$, $P < 0.001$), and political unacceptability ($b=1.66$, 95% CI=[0.96,2.35], OR=5.24, $z=4.69$, $P < 0.001$). There was no significant main effect of party identification on disliking ($b=-0.04$, 95% CI=[-0.37,0.30], OR=0.96, $z=-0.22$, $P=0.83$), opposition ($b=-0.11$, 95% CI=[-0.44,0.23], OR=0.90, $z=-0.61$, $P=0.55$), or political unacceptability ($b=-0.08$, 95% CI=[-0.42,0.26], OR=0.92, $z=-0.47$, $P=0.64$). Thus, the more negative (and therefore inaccurate)

participants' GMPs were, the more likely they were to believe the out-group is motivated by obstructionism. See Figure 1.3 for visualization of raw data and Pearson correlations.

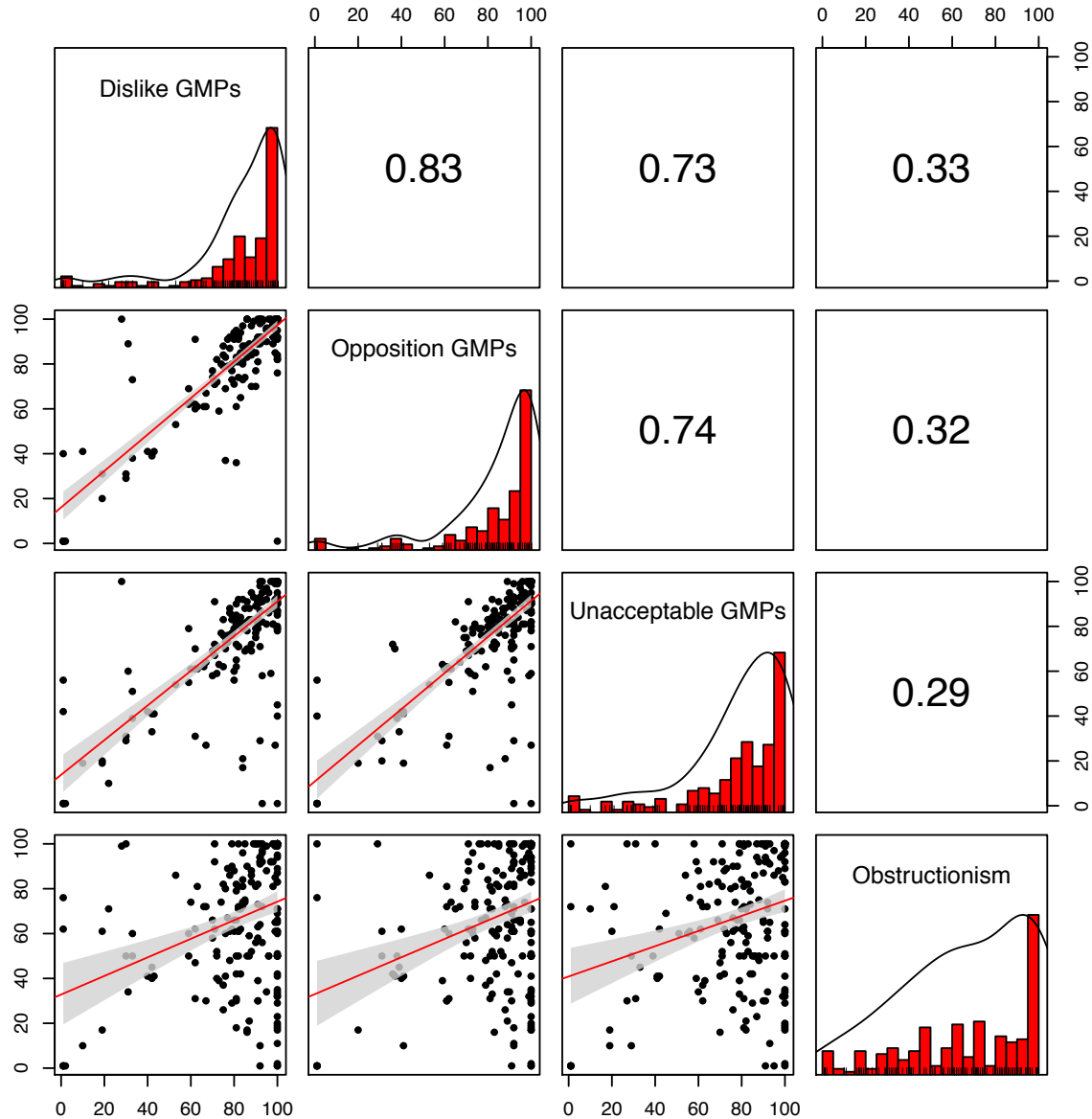


Figure 1.3: Distributions, Pearson correlations, and scatterplots for the three GMP ratings and beliefs about out-group obstructionism in Study 5. Sample size, $N=212$ (collected via Mechanical Turk). Scatterplot lines are linear regression lines, shaded area around lines are 95% confidence intervals. Correlations: Disliking – Opposition ($r=0.83$, 95% CI=[0.79,0.87], $t(208)=21.73$, $P < 0.001$), Disliking – Unacceptable ($r=0.73$, 95% CI=[0.66,0.79], $t(210)=15.50$, $P < 0.001$), Disliking – Obstructionism ($r=0.33$, 95% CI=[0.20,0.45], $t(210)=5.08$, $P < 0.001$), Unacceptable – Opposition ($r=0.74$, 95% CI=[0.68,0.80], $t(208)=16.02$, $P < 0.001$), Unacceptable – Obstructionism ($r=0.29$, 95% CI=[0.16,0.40], $t(210)=4.32$, $P < 0.001$), and Obstructionism – Opposition ($r=0.32$, 95% CI=[0.19,0.43], $t(208)=4.80$, $P < 0.001$).

0.001). These data indicate a positive linear association between pessimistic GMPs and the belief that the out-group is purposefully obstructionist.

Experiment 6 (N=1122) sought to reduce the perception that the out-group is motivated by obstructionism by utilizing a preregistered intervention (see OSF: <https://osf.io/jhnsb>). Building upon Study 5's design, after participants provided their three GMP ratings in response to one of the five political scenarios, participants were randomly assigned, between-subjects, to one of three conditions before reporting their perceived out-group obstructionism: the control (N=396), "truth intervention" (N=358), or "hypocrisy prevention intervention" (N=368) conditions. In the control condition participants were simply reminded of the GMP ratings they had provided on the previous page (i.e., no new information). In the truth intervention, participants were provided with the information from the control condition plus the true value for their out-group's actual perceptions (the mean of the representative sample responses from Experiment 4) for that same scenario. This allowed participants to see the (in)accuracy of their GMPs. Recall that in Experiment 4 we also found that participants inaccurately believed their in-group would react less negatively than their out-group to the same behavior. Therefore, in the hypocrisy prevention intervention participants received all the information in the truth intervention while also receiving the exact true values for their in-group's actual perceptions (also drawn from Experiment 4), for the same scenario. As such, the hypocrisy intervention additionally prevented participants from anchoring on an inaccurate belief that the in-group's negativity would still be lower than the out-group's in the same scenario. This allowed us to test whether there was an added benefit to highlighting participants' (in)accuracy regarding the extent to which their in-group and out-group were similar in their actual perceptions.

As hypothesized, participants who were assigned to the truth intervention condition had lower ratings of out-group obstructionism than did the control group ($b=-4.08$, 95% CI=[-7.67,-

0.48], $\beta=-0.155$, $t(1114)=-2.22$, $P=0.027$). Those assigned to the hypocrisy prevention intervention also had lower obstructionism ratings relative to control ($b=-4.64$, 95% CI=[-8.22,-1.08], $\beta=-0.177$, $t(1114)=-2.55$, $P=0.011$). However, post-hoc pairwise comparisons indicated no statistically significant difference in obstructionism between the hypocrisy prevention and truth interventions ($b=-0.57$, 95% CI=[-4.96,3.82], $t(1115)=-0.304$, $P=0.95$), suggesting the hypocrisy prevention intervention provided no additional benefit above the truth intervention. There was also a main effect of party identification on obstructionism, with Democrats rating Republicans as higher on obstructionism than Republicans rated Democrats ($b=-3.84$, 95% CI=[-6.88,-0.79], $\beta=-0.146$, $t(1114)=-2.47$, $P=0.014$); however, further analysis indicated no statistically significant party by condition interaction for either the truth intervention ($b=4.44$, 95% CI=[-2.97,11.88], $t(1113)=1.17$, $P=0.24$), or hypocrisy intervention ($b=0.83$, 95% CI=[-6.59,8.26], $t(1112)=0.22$, $P=0.83$). In other words, the interventions were not more effective at reducing negative motive attributions among one party relative to the other.

Further analysis revealed statistically significant interactions of condition on GMP inaccuracy (operationalized as the mean difference between participants' GMPs and the true values, such that higher values = more inaccurate and negative). We found that GMP inaccuracy moderated the effectiveness of the hypocrisy prevention intervention ($b=-0.17$, 95% CI=[-0.33,-0.01], $\beta=-0.144$, $t(1112)=-2.09$, $P=0.037$), and truth intervention ($b=-0.27$, 95% CI=[-0.43,-0.12], $\beta=-0.23$, $t(1113)=-3.39$, $P < 0.001$), relative to control. In other words, the interventions were more effective at reducing obstructionism for participants whose GMPs were relatively less accurate and more negative. There was also a linear association between inaccuracy and perceived obstructionism ($b=0.44$, 95% CI=[0.32,0.56], $\beta=0.37$, $t(1114)=7.33$, $P < 0.001$), replicating the finding from Study 5. See Figure 1.4 (continued) for visualization of the effect of

the interventions at one standard deviation above and below the mean of accuracy (see Supplementary Figure 4 for raw data distributions). As an exploratory measure, we followed up with participants one week after they completed Experiment 6 to see if the effect of the intervention persisted over time. We had a 73% response rate, but found no credible evidence for a continued effect of the intervention on a rating of general out-group obstructionism (see “Supplemental Experiment B”).

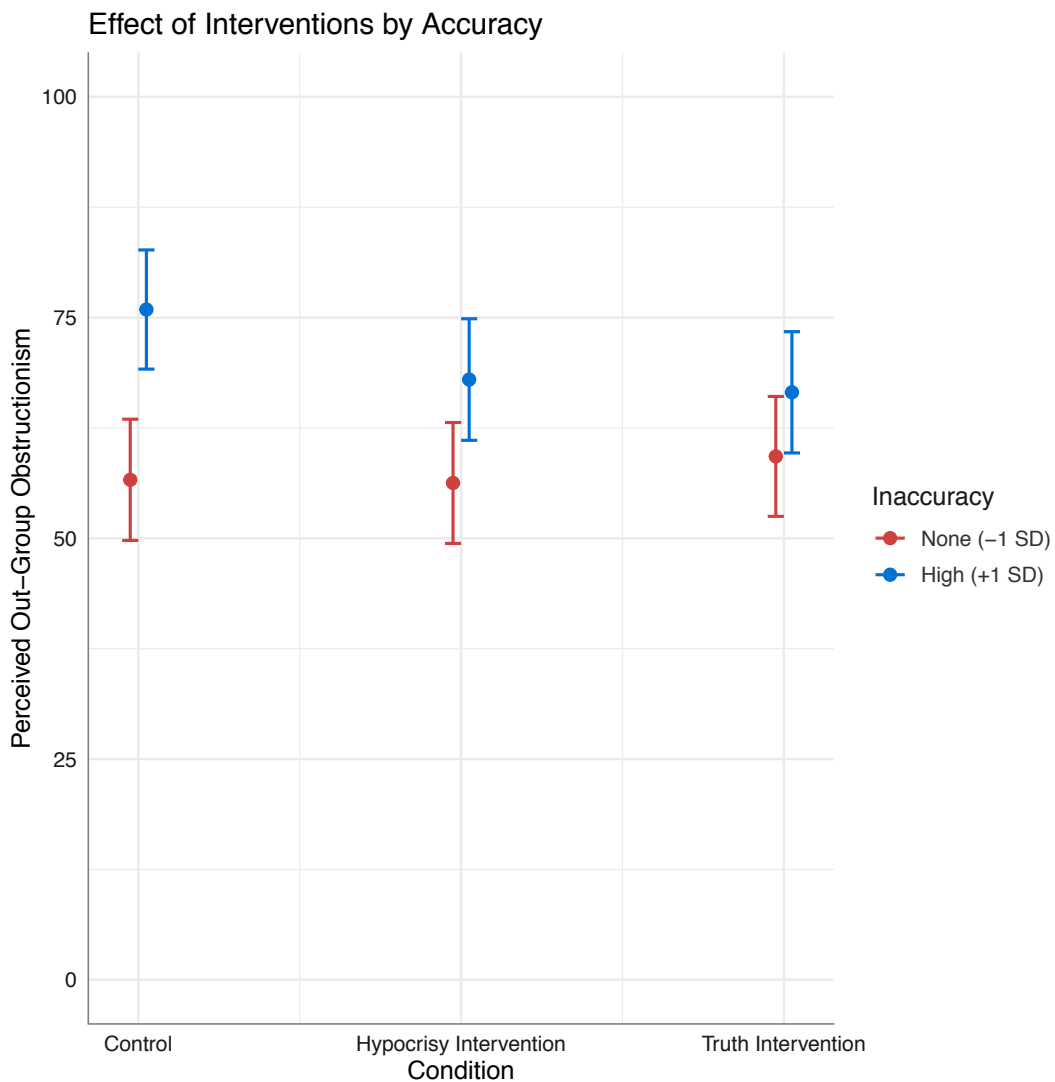


Figure 1.4: Effect of condition on obstructionism, by accuracy, in Experiment 6. Sample size, $N=1122$ (collected via Mechanical Turk). By Condition: Control=396, Hypocrisy Intervention=368, Truth Intervention=358. GMP inaccuracy moderated the effectiveness of the hypocrisy prevention intervention ($b=-0.17$, 95% CI=[-0.33,-0.01], $\beta=-0.144$, $t(1,112)=-2.09$, $P=0.037$), and truth intervention ($b=-0.27$,

95% CI=[-0.43,-0.12], $\beta=-0.23$, $t(1,113)=-3.39$, $P < 0.001$) at reducing obstructionism. In other words, the interventions were more effective at reducing obstructionism for participants whose GMPs were relatively more inaccurate and negative. Here inaccuracy is plotted at one standard deviation above and below the mean inaccuracy ($M=22$, $SD=22$). -1 SD equals an inaccuracy of zero, meaning that the participant was on average perfectly accurate in their GMPs. +1 SD equals an inaccuracy of 44, meaning that the participant on average overestimated out-group negativity by 44 points (on a 100-point scale). Bars are 95% confidence intervals.

The results of Experiment 6 provided support for the hypothesis that negative motivational attributions towards the out-group, such as obstructionism, were driven in part by inaccurate beliefs regarding how negatively the out-group perceived the collective behavior of one's in-group. They also suggest that simply providing individuals with concrete information regarding their inaccurate, and overly negative, GMPs can help reduce downstream negative attributions towards the out-group. However, we found no credible evidence that the hypocrisy prevention intervention provided additional benefit above the truth intervention, which suggests that participants were not anchoring on inaccurate beliefs about how the in-group would react to the same behavior. Given the central role motive attributions play in intergroup relations (Miller & Nelson, 2002; Reeder et al., 2004), our findings highlight a potential avenue for future attempts at reducing intergroup hostility and conflict, and an avenue for further understanding the antecedents of negative and inaccurate motive attributions.

Discussion

Across seven experiments and one survey we found that group meta-perceptions were consistently inaccurate and negatively biased across a variety of competitive intergroup contexts, scenarios, and participant samples. Theoretically, our findings of negative and inaccurate GMPs across multiple intergroup domains—even in the absence of group labels as in the control condition of Experiment 1—parallel research on the interindividual-intergroup discontinuity effect (IIDE), which demonstrates that intergroup interactions are more hostile and competitive than interindividual interactions (Insko et al., 1990; Wildschut et al., 2003). Importantly, the

IIDE is observed both in actual behavior and in expectations of behavior, in that people expect future intergroup interactions to be more hostile than interpersonal interaction (Pemberton et al., 1996). If people assume that intergroup interactions are going to be more hostile, this may partially explain why GMPs are overly negative and associated with negative motive attributions, although it does not explain why GMPs are so inaccurate. Similarly, while recent evidence suggests that perceptions of political party polarization in the US have become more negative and inaccurate over the past four decades (Enders & Armaly, 2018; Westfall et al., 2015), this does not explain inaccurate GMPs in the domain of gender, why there is no evidence for GMP inaccuracy in cooperative political contexts, and why there is no evidence that inaccurate GMPs vary across the scenario content or party of the perceiver.

Several limitations in these experiments highlight fruitful avenues for future research. One assumption embedded in these studies is that actual perceptions represent ground truth. An alternative source of GMP inaccuracy may be actual perceivers downplaying their reactions to these events. For example, in Experiment 2, men might have been underreporting their dissatisfaction with losing resources, which would make women's GMPs look more inaccurate than they are. Furthermore, the use of random-probability sampling would be superior to the quota-matching methods we used in Experiment 4 for estimating the true population 'actual-perceptions' of our scenarios. Second, we did not measure confidence in participants' own judgments, which should be related to GMP (in)accuracy as it is in other meta-perception research (Carlson et al., 2010). Third, we found no statistically significant effect of our intervention on negative motive attributions one week after it was administered, though we hasten to note that we specifically designed our intervention to minimize the likelihood that our results were driven by demand effects. Furthermore, the attrition-rate of participants meant our

follow-up measurement one week later was likely underpowered. Future research should vary the strength and nature of any such interventions in order to understand better which qualities provide more (if any) benefit over time.

Conceptually, future research ought to examine the relationship between GMPs and other second-order judgments in intergroup contexts. Here we operationalized GMPs as judgments regarding out-group members' reactions to collective in-group behaviors, but GMPs can be measured along many features, including attitude (Stern & Kleiman, 2015) and trait (Stroessner & Dweck, 2015) attributions (i.e., "how they see us"), dehumanization (Kteily et al., 2016) (i.e., "how human they think we are"), judgments of intent (Ames & Fiske, 2015), even group emotions (Goldenberg et al., 2014). Understanding how GMPs across these judgments relate to, and are distinct from, one another will be critical in building theory around the dynamics of and outcomes associated with GMPs in intergroup contexts. Lastly, future work should also seek to take advantage of current events as they are unfolding in order to see how inaccuracies in GMP are shaped during real world events related to issues with which people are very familiar.

Our findings highlight a consistent, pernicious inaccuracy in social perception, along with how these inaccurate perceptions relate to negative attributions towards out-groups. More broadly, inaccurate and overly negative GMPs exist across multiple competitive intergroup contexts, and we find no evidence they differ across the political spectrum. This suggests that there may be many domains of intergroup interaction where inaccurate GMPs could potentially diminish the likelihood of cooperation and instead exacerbate the possibility of conflict. However, our findings also highlight a straight-forward manner in which simply informing individuals of their inaccurate beliefs can reduce these negative attributions.

Methods

All studies were approved by Harvard University's Institutional Review Board, and all participants gave their informed consent before participating. All participants, except those in Experiment 4, were collected on Amazon's Mechanical Turk platform ("Mturk"), and were located in the United States. Participants in Experiment 4 were collected through Qualtrics Survey Panels, and the sample was quota-matched to US census data distributions of the following variables in the general population: age, gender, ethnicity, education, and income (see supplemental materials for demographic breakdown and quotas). All surveys were administered via the Qualtrics survey platform.

Participants. Samples from Experiments 1, 3, 4, 6 and Study 5 consist of self-identified Republicans and Democrats, and the sample of Experiment 2 consists of self-identified men and women. Experiment 1 (N=408) and Experiment 2 (N=286) had sample sizes of 170 per condition determined a priori with the goal of attaining 144 per condition after excluding participants who failed comprehension checks (see Exclusions section below). An a priori power analysis indicated that 144 per condition was necessary to detect a small effect size of $f=0.15$ with 80% power within a three condition between-subjects ANOVA framework. Expecting to observe a reduced effect size in Experiment 3 (N=499) relative to Experiment 1, we increased the sample size to a target of 275 per condition, and collected 675 in the hopes of reaching 550 after exclusions. We did not conduct a formal power analysis for Experiment 3. Experiment 4 had a preregistered sample size of N=500 (selected via a priori power analysis to detect standardized $b = 0.20$ with 80% power; see preregistration for details); Qualtrics purposefully oversampled to ensure a minimum of 500 quality responses (hence final N=536). For Study 5 (N=212) we selected an a priori sample size of N=300, with the goal of attaining approximately N=250 after

exclusions, the sample size at which small correlations stabilize (Schönbrodt & Perugini, 2013). Experiment 6 (N=1122) had a preregistered sample size of N=1510, in the hopes of obtained 1260 after exclusions (selected via a priori power analysis to detect standardized $b = 0.20$ with 80% power; see preregistration for details)

Exclusions: In Experiment 1 we removed 12 responses due to three separate participants taking the study multiple times (all their responses were removed). A further 89 participants failed the comprehension check, and one participant was excluded for not completing the dependent variable ratings, leaving a final N=408 (mean age (M_{age}) = 35.2, 239 Women). In Experiment 2 we removed two responses due to one participant completing the study twice, another response due to a participant not providing their gender identity, and 56 participants who failed the comprehension check, leaving a final N=286 ($M_{age} = 36.2$, 156 Women). In Experiment 3, 165 participants failed the comprehension check, and 12 responses were removed due to duplicate IP addresses, leaving a final N=499 ($M_{age} = 35.1$, 293 Women). In Experiment 4, 364 participants failed the comprehension check, and the Qualtrics manager continued collecting data until 536 participants (273 Women, Age Brackets: 165 in ages 18-34, 189 in ages 35-54, 182 in ages 55+), who met our demographic quotas, completed the study. In Study 5, 86 participants failed the comprehension check, and two participants were removed for not completing the dependent variable rating, leaving a final N=212 ($M_{age} = 35.89$, 120 Women). In Experiment 6, 349 participants failed the comprehension check, and 26 responses were removed due to duplicate Mturk ID or IP addresses, leaving a final N=1122 ($M_{age} = 35.1$, 642 Women). We did not weigh Mturk samples by political party or gender because we were interested in in-group versus out-group dynamics, not the difference between, for example, Democrats and Republicans. In Experiment 4 we quota-matched to a 50/50 split of Democrats and Republicans.

Self-identified Independents were allowed to complete all studies (except Experiment 4), but were excluded from all analyses a priori.

Compensation: Experiments 1, 2, 3, and Study 5 paid \$0.10 and were advertised as taking one minute. Experiment 4 was advertised as taking 9 minutes (itself 4 minutes, but it was bundled with a separate 5-minute study which always came after Experiment 4), and participants were paid a preset amount of credit via Qualtrics Panel's internal payment system. Experiment 6 paid \$0.15 and was advertised as taking 60-90 seconds.

Procedure: We randomly assigned participants to condition and scenario (in Experiment 4 scenario order) across all the experiments and studies. Across scenarios, we also randomized the order of the dependent variable items (e.g. disliking, opposition). All randomization was facilitated through Qualtrics' randomization functions. The surveys were programmed to pipe the appropriate out-group and/or in-group labels into the scenarios and dependent variables ratings based on the participants' self-reported group affiliation. All dependent variables across all studies appeared as sliding scales with end-labels and tick-marks, but no visible numbers (except for the ratings in Experiment 6, in which a numeric value (1-100) appeared next to the slider when participants provided a response). Across all experiments and studies, except Experiment 4, excluded participants received full compensation.

Analyses: We analyzed Experiments 1-4 and Study 5 using mixed-effects beta-regressions (glmmTMB (Brooks et al., 2017) package, v 0.2.3) in R (v 3.6.1) and Experiment 6 using linear mixed-effects modeling (lmerTest (Kuznetsova et al., 2017) R package, v 3.1-0). All post-hoc tests utilized the Tukey method for *P*-value adjustment and were conducted with the emmeans

(Lenth, 2019) R package (v. 1.4). We used beta-regressions for Experiments 1-4 and Study 5 due to the highly skewed GMP response data, and transformed the data for the beta-regressions using established formulas (Smithson & Verkuilen, 2006). As a robustness check we performed all non-preregistered beta-regression analyses (Experiments 1, 2, 3, and Study 5) using linear mixed effects modeling via the lmerTest R package: none of our results changed meaningfully. For Experiments 1, 3, 4, Study 5, and the main effects of Experiment 6, we report the results from models that include only the main effects because there were never any significant interactions among the fixed effects; furthermore, the saturated models including fixed effects and the corresponding interactions did not improve model fits. Results for Experiment 2 are from the saturated models, and while we report the interaction of accuracy on condition in Experiment 6, we never find an interaction with party identification and do not report those saturated models. Across Experiments 1-4 we regressed the relevant dependent variable rating (dislike, opposition, political/social unacceptability) onto fixed effects for condition and the relevant group variable (“party accuracy” in Experiment 1, 3 and 4, “gender accuracy” in Experiment 2), a random effect with random intercepts for scenario (along with an random effect with random intercepts for participant in Experiment 4, due to the repeated measures), and in Experiment 2 an interaction term for the condition by group interaction. In Study 5 we regressed obstructionism onto each GMP item separately, including a fixed effect for party and a random effect with random intercepts for scenario. In Experiment 6 we regressed obstructionism onto condition including a fixed effect for party and a random effect with random intercepts for scenario, then replaced the fixed effect for party with the interaction of accuracy with condition. All tests were two-sided. Data analyses were not performed blind to the conditions of the experiments and studies. Figures

were created using the R packages ggstatsplot (Patil & Powell, 2018) (v. 0.0.12), sjPlot (Lüdecke, 2019) (v. 2.7.0), and psych (Revelle, 2018) (v. 1.8.12).

Experiments 4 and 6 were preregistered. Experiment 4 was preregistered on February 26th, 2019 and can be found here: <https://osf.io/atck5>. Experiment 6 was preregistered on March 19th, 2019 and can be found here: <https://osf.io/jhnsb>. No analyses deviate from the preregistrations.

Data Availability

All data that supported the findings of this study are publicly available in CSV format on the Open Science Framework: <https://osf.io/zhysa/>

Code Availability

All analyses reported in this study used the statistical software R (v 3.6.1). All R files are publicly available on the Open Science Framework: <https://osf.io/zhysa/>

Acknowledgments

Work on this project by MC was supported by a National Science Foundation Award (BCS-1551559). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. JL received no specific funding for this work. We thank members of the Harvard Intergroup Neuroscience Lab, Sidanius Lab, and attendees to the 2018 East Coast Doctoral Conference for their helpful comments, Z. Ingbretsen and N. Hunt for help with data collection, and I. Zahn and S. Worthington for statistical assistance.

Author Contributions

J.L. and M.C. designed all experiments and wrote the manuscript. J.L. completed data collection and analysis under the supervision of M.C.

Competing Interests

The authors declare no competing interests

Chapter 2

Moral prospection: Cognitive bias and the failure to predict moral backlash toward an organization

Jeffrey Lees

Abstract

This theoretical paper elucidates a novel psychological phenomenon that drives corporate social irresponsibility and stakeholder backlash: *moral prospection*. The theory of moral prospection proposes that when confronted by a decision with clear moral implications organizational leaders generate moral meta-perceptions, their estimations of the moral attributions stakeholders will make of the organization's behavior, and that these meta-perceptions are inaccurate due to a cognitive bias in this simulation. The theory argues that a *corporate personhood bias* exists between how leaders and stakeholders perceive organizations, and that leaders fail to take this discontinuity into account when simulating stakeholders' moral attributions toward their organization. Prospectors therefore generate overly positive moral meta-perceptions which produce a state of *moral overconfidence*, leading to behaviors which engender moral backlash from stakeholders. The attributional domain of intent is identified as the primary one of prospective inaccuracy, which then drives inaccuracies in meta-judgments of harm caused, organizational culpability, and non-complicity of affected parties. The theory of moral prospection integrates insights from moral psychology with corporate social irresponsibility scholarship to better understand a novel cognitive phenomenon which contributes to corporate scandal and moral backlash toward organizations.

On March 15th, 2015 the popular coffee company Starbucks announced their “Race Together” campaign. Described as “an opportunity to begin to re-examine how we can create a more empathetic and inclusive society – one conversation at a time” (Starbucks Corporation, 2015a), the campaign centered on baristas attempting to actively engage US customers in discussions about how to improve race relations in the United States while the customer was in the store. In preparation for the campaign Starbucks took out full-page advertisements in The New York Times and USA Today. Despite the intent to “stimulate conversation, empathy and compassion” the campaign was met with intense public and media backlash. Criticism of the campaign included questioning Starbucks’ moral authority to even engage in such an effort, highlighting perceived company hypocrisy on matters of race relations, anger that the burden of the campaign fell almost entirely on low-level employees, and challenging the company’s claims to sincere intent (Logan, 2016). In the face of sustained public backlash the campaign was cancelled by then CEO Howard Schultz only seven days after its announcement (Ember, 2015; Starbucks Corporation, 2015b).

Scholars in the bounded ethicality tradition have long noted a host of cognitive biases that lead to “ethical blind spots” in our judgment and behaviors (Bazerman & Gino, 2012; Sezer et al., 2015; Tenbrunsel et al., 2010). This literature provides a wealth of evidence that we are often “blind” to the moral nature of our decisions, and that this blindness accounts for many behaviors which are considered immoral by third-party observers (and often the decision-maker themselves, ex post). However, in the case of “Race Together,” Starbucks was well aware of the moral content of their behavior, and in fact made explicit moral appeals, both to the public and internally within the company, when motivating the campaign (Carr, 2015; Logan, 2016; Starbucks Corporation, 2015a). The moral backlash to the Race Together campaign does not

seem to have been a result of moral blindness on the part of Starbucks or their then CEO Howard Schultz, per se. Rather, Starbucks' error was in their belief that Race Together would be perceived in a morally positive light by the public, just as they themselves perceived the campaign. This was a failure to *accurately predict* how others would perceive their moral behavior, not a failure to realize their behavior contained moral content.

This theoretical paper seeks to understand this phenomenon of inaccurately estimating how one's organization will be perceived in the moral domain. In this paper I integrate insights from behavioral ethics and moral psychology with scholarship on corporate social irresponsibility (CSIR) to better understand how people morally perceive organizations, how leaders *think* people will morally perceive their organization, the psychological mechanisms underlying why such estimates may be inaccurate, and how such inaccuracies ultimately lead to organizational decisions that evoke moral backlash. The theory centers on the cognitive process of moral prospection (Railton, 2016), defined broadly as the act of attempting to estimate how others would perceive one's potential moral behaviors. Here I focus specifically on how moral prospection functions when attempting to estimate how others would perceive the moral behavior of *one's organization*.

The theory argues, based on recent psychological scholarship, that anthropomorphic judgments toward organizations (i.e. how much one believes an organization possesses human-like mental capacities) are deterministic of moral judgments toward organizations (Jago & Laurin, 2017; Rai & Diermeier, 2015; Schein & Gray, 2018). When organizational leaders are confronted with a moral decision I argue, based on recent work in moral philosophy (Railton, 2016), they attempt to cognitively simulate how stakeholders will morally perceive and react toward a potential decision or strategy. In doing so, prospectors are subject to a cognitive bias I

call the *corporate personhood bias* (Kervyn et al., 2012; Knobe & Prinz, 2008; Rai & Diermeier, 2015), where they fail to realize that stakeholders anthropomorphize the organization *less* than do the prospectors. This failure to accurately simulate the anthropic perceptions (or the lack thereof) of stakeholders will lead to inaccurate and overly positive moral meta-perceptions: the prospector's final judgment as to how the organization will be morally perceived by stakeholders. Specifically, moral prospectors will be inaccurate in the attributional domain of intent (Ames & Fiske, 2015; Cushman, 2015; Kervyn et al., 2012), which then drives inaccuracies in meta-judgments of harm caused, organizational culpability, and non-complicity of affected parties (Lange & Washburn, 2012). Through a process akin to moral learning (Cushman et al., 2017; Railton, 2017) these inaccurate moral meta-perceptions engender a state of *moral overconfidence*, where the inaccurate meta-perceptions are used by the prospector to overcome uncertainty regarding the moral rightness of a given decision (Reynolds et al., 2012) and codify the prospector's belief that a particular decision is indeed reflective of their own moral values. Moral overconfidence then leads to a decision that evokes moral backlash from stakeholders, and drives organizational leaders to respond to such backlash with defensiveness and incredulity rather than contrition.

The theory of moral prospection operates within a specific organizational context. First, the theory of moral prospection applies to situations where an organizational leader is confronted with a decision which has moral content that is *clear to the decision-maker*. That is to say, said decision-maker is not subject to, or has overcome, the host of cognitive biases which drive moral blindness (Sezer et al., 2015). For ease of reference, I hereafter refer to such a decision-context as one of "moral quandary." Second, the theory of moral prospection pertains to organizational decision-makers in for-profit organizations. The cognitive bias highlighted by the theory of

moral prospection, the corporate personhood bias, pertains explicitly to the discontinuity in how those within and those outside typical for-profit companies perceive such companies. How the process of moral prospection might function for individuals inside non-profit organizations is elaborated on in the Discussion section.

The theory of moral prospection makes three important contributions to management scholarship. First, it contributes to traditions in behavioral ethics (Bazerman & Gino, 2012) and the micro-foundations of CSIR (Gond et al., 2017) by highlighting a novel moral phenomenon that causes leaders to engage in irresponsible behavior, and has yet to be explored systematically by either literature. Second, it contributes to research on the psychological drivers of CSIR judgments (Lange & Washburn, 2012; Murphy & Schlegelmilch, 2013) by integrating recent scholarship in moral psychology on anthropomorphization (Epley et al., 2007; Gray et al., 2012) with existing CSIR theory. Third, by identifying the corporate personhood bias the theory provides a clear avenue for future research attempting to develop interventions that can reduce immoral behavior in organizations (Epley & Tannenbaum, 2017; Zhang et al., 2014).

Psychology and Corporate Social Irresponsibility

While a large body of scholarship has detailed the psychological processes underlying engagement with and reactions to corporate social responsibility (CSR) (Glavas, 2016), relatively little research has focused on the psychology of corporate social *irresponsibility* (CSIR) (Gond et al., 2017; Murphy & Schlegelmilch, 2013). Furthermore, while stakeholder management has long been central to CSR theory (Swanson, 1995; Wood, 1991) the majority of micro-CSR research has focused on how employees perceive and react to CSR initiatives (Rupp & Mallory, 2015). This literature has detailed how employees' values (Giacalone et al., 2008; Humphreys & Brown, 2008), emotions (Grappi et al., 2013) and trust (De Roeck & Delobbe,

2012; Vlachos et al., 2010) drive CSR reactions and affect employee identification with the organization (De Roeck & Maon, 2016; Farooq et al., 2017; Glavas & Godwin, 2013). Similarly, research on the antecedents on leader/manager CSR engagement has highlighted the role of perceived stakeholder pressure (Wang et al., 2015), organizational ethics codes (Stevens et al., 2005), prescriptive vs. proscriptive mindsets (Stahl & de Luque, 2014), emotions (Crilly et al., 2008), and moral beliefs (Hibbert & Cunliffe, 2015; Weaver et al., 1999) on when and how leaders choose to engage in CSR. However only a few papers have examined the psychological antecedents of CSIR. For example, leaders' need for power can drive CSIR (Pearce & Manz, 2011), but shared leadership can mitigate this relationship (Fabrizi et al., 2014).

Research has also detailed how some stakeholders react towards CSIR. Consumers are often incredulous towards corporate CSR initiatives (Skarmeas & Leonidou, 2013), and instances of CSIR can evoke strong moral emotions in consumers, who are willing to punish organizations and brands they perceive as irresponsible (Grappi et al., 2013; Palazzo & Basu, 2007; Sweetin et al., 2013). Investors too will punish organizations for CSR initiatives they perceive as negative, even when those initiatives benefit other stakeholders (Groening & Kanuri, 2013).

Lange & Washburn (2012) leverage research on attribution theory to develop theory surrounding when an organization's behavior will be perceived as irresponsible, and they highlight three central attributional dimensions that all contribute to the perception of CSIR. The first domain is observers' assessment of effect undesirability, which is based on threat avoidance, moral impulses, and norms of moral behavior. The second dimension is observers' assessment of corporate culpability, based on inferences of causality and judgments of moral responsibility. Lastly, observers' make assessments of affected party non-complicity, based on judgments of

power to prevent the effect and of foresight. Lange & Washburn's (2012) model also explicates the role observer identity, firm and effect characteristics, and framing have on these attributional dimensions.

This theoretical framework, while not drawing directly from moral psychology, shares similar predictions with recent work on the cognitive mechanisms underlying moral judgments, and how those mechanisms relate to the process of anthropomorphization toward organizations. Integrating insights from moral psychology, specifically the theory of dyadic morality (Gray et al., 2012; Schein & Gray, 2018) with Lange & Washburn's (2012) model of CSIR attributions allows for a robust understanding of how stakeholders make moral judgments of organizations, and what psychological processes drive those judgments. Such an understanding of how people perceive organizations is necessary in order for the theory of moral prospection to assess how leaders may inaccurately estimate such judgments.

The theory of dyadic morality (Gray et al., 2012; Schein & Gray, 2018) argues that all moral judgments are fundamentally constituted by a single cognitive process called "mind perception" (Epley & Waytz, 2010). Mind perception is the process of determining whether an entity is capable of certain mental states, and is referred to as "anthropomorphization" when the target of mind perception is non-human (Epley et al., 2007). Importantly, mind perception falls along two dimensions, "agency" and "experience" (Gray et al., 2007). Agency encompass mental capacity attributions such as self-control, intelligence, and intentionality, whereas experiential attributions encompass the capacity for emotions, feeling pain and pleasure, and desire. Dyadic morality theory argues that all moral judgments are simulated in the mind as an interaction between two (hence "dyadic") entities, a "moral agent" and "moral patient." The moral agent is the transgressor and the patient the victim, and the extent to which we judge an action as wrong

depends primarily on mental state attributions toward the moral agent (i.e. are they capable of intentionally causing harm), and toward the moral patient (i.e. are they capable of experiencing suffering). For debates over the merits of dyadic morality theory see: Bauman, Wisneski, & Skitka (2012); Dillon & Cushman (2012); Graham (2015); Gray & Keeney (2015); Monroe, Guglielmo, & Malle (2012).

While this framework has never been formally integrated with CSIR scholarship, the cognitive mechanisms highlighted by dyadic morality theory, and parallel work across moral psychology, cohere with the moral attributional domains that determine judgments of organizational irresponsibility (Lange & Washburn, 2012). The first dimension Lange & Washburn (2012) detail is that of effect undesirability, which aligns closely to judgments of harm in the dyadic morality framework (Gray et al., 2014; Schein & Gray, 2016). The second dimension is corporate culpability, which encompasses both judgments of causality, which are studied directly in moral psychology (Cushman, 2008, 2013), and judgments of blame, which are encompassed by the dyadic morality framework (Gray & Wegner, 2011; Schein & Gray, 2014). The third dimension is affected party non-complicity, which aligns closely to judgments of victimhood status in dyadic morality theory (Gray & Wegner, 2009, 2011). While the dimensions of CSIR judgments proposed by Lange & Washburn (2012) adhere closely to empirical research in moral psychology, here I will argue that such judgments are directly predicated on first-order judgments of *intent*. A growing body of psychological scholarship suggests that judgments of intent in moral contexts drive judgments of harm (Ames & Fiske, 2015; Chakroff et al., 2016; Gray, 2012), judgments of causality (Cushman, 2015; Pizarro et al., 2003) judgments of blame (Cushman, 2008; Schein & Gray, 2014; Young & Saxe, 2011), and judgments of victimhood (Gray & Wegner, 2009; Niemi & Young, 2014, 2016). Nonetheless,

most research in moral psychology focuses on perceptions of human targets. To understand how mind perception toward organizations drives perceptions of intent, which then affect CSIR judgments, we must first understand how mind perception functions when perceiving organizations.

Scholars studying mind perception have begun to examine how it functions across non-human targets (“anthropomorphization”), such as autonomous vehicles (Waytz et al., 2014), brands (Puzakova et al., 2013), social causes (Ahn et al., 2014) and algorithms (Jago, 2019). Research examining anthropomorphization toward corporations has found that they are “cyborgs,” in that they are attributed equal levels of agency as humans, but significantly less experience (Kervyn et al., 2012; Knobe & Prinz, 2008; Rai & Diermeier, 2015). Put plainly, corporations are perceived as equally capable of intelligence, competence, and intention as humans are, but less capable of having emotions, empathy, or feeling pain and pleasure relative to humans. This means that while companies are perceived as capable of having and acting on discrete intentions, they are not seen as capable of altruistic, emotion-based, and sympathetic intentions (Kervyn et al., 2012). Critically, Rai & Deirmeier (2015) found that these high agency/low experience perceptions were held by a general population sample, but when they asked a sample of senior executives about hypothetical companies these executives fully anthropomorphized them, in that senior executives attributed these companies a level of agency *and experience* equal to that of human targets. This finding constitutes the *corporate personhood bias*, the discontinuity between how organizational leaders and those outside the organization (e.g. most stakeholders) anthropomorphize organizations. Namely, that those who strongly identify as members of an organization will fully anthropomorphize their organization (Epley et al., 2007; Sluss & Ashforth, 2008), while external stakeholders will perceive the organization as

agentic but not capable of emotions (Rai & Diermeier, 2015). The complexities surrounding whether employees anthropomorphize their organization to the same extent as leaders, and variance as such across stakeholders in general, is detailed in the Discussion section.

The Moral Prospecion Model

The theory of moral prospecion highlights the processes by which organizational leaders inaccurately estimate stakeholders' perceptions of and reactions to the organization's potential behavior, and how these inaccurate meta-perceptions can lead to decisions that engender moral backlash from stakeholders. Moral prospectors, when confronted with a moral quandary, begin by engaging in an intuitive, iterative, future-oriented cognitive simulation (Seligman et al., 2013) where they attempt to imagine stakeholders' perceptions of and reactions toward the moral decision at hand. Because this simulation requires estimating multiple observers perceiving a myriad set of possible organizational decisions, moral prospecion is largely, but not entirely, a "System 1" cognitive task governed by automatic processing, affect-guided intuition, and heuristic-based judgments (Gaesser & Schacter, 2014; Gaesser et al., 2017; Gilbert & Wilson, 2007; Kahneman, 2011; Reynolds et al., 2010). Due to the intuitive nature of moral prospecion, prospectors fail to anticipate the discontinuity in anthropomorphization between themselves and the stakeholders they are simulating (the corporate personhood bias). This failure in accurate simulation leads to inaccurate, and overly positive, moral meta-perceptions: the final estimations of how stakeholders will perceive and react to the organization's potential behaviors.

These meta-perceptions will be inaccurate across specific attributional domains. The foremost domain of meta-perceptive inaccuracy will be that of perceived sympathetic intent. Prospectors will inaccurately simulate organizational outsiders as perceiving the organization as capable of decisions driven by emotional, altruistic, or sympathetic motives. This inaccurate

estimation of intent attributions will drive further inaccuracies in other meta-perceptive judgments directly tied to intent attributions. Specifically, prospectors will underestimate the perceived level of harm caused by the organization's behavior, underestimate the perceived level of causality between the organization's behavior and the outcomes engendered by that behavior, underestimate judgments of organizational blameworthiness, and overestimate perceptions of affected party complicity.

These inaccurate meta-perceptions serve to create a psychological state of moral overconfidence. Here moral overconfidence is not merely defined as the possession of inaccurate moral meta-perception, but rather entails two further components. First, it is defined as over-precision (Moore et al., 2010) in the veracity of meta-perceptive judgments, such that prospectors have undue confidence that their meta-perceptions are accurate. Second, moral overconfidence is defined by a process akin to moral learning (Cushman et al., 2017), where the generated moral meta-perceptions are used by the prospector to update their own moral preferences (Railton, 2017; Schacter et al., 2007). As prospectors come to believe others will see a particular decision as moral, they themselves begin to personally believe said decision is moral. Moral overconfidence then drives a moral decision that will engender unexpected moral backlash from stakeholders, as the leaders have inaccurate and overly positive expectations regarding the reaction of stakeholders and newly formed moral preferences not aligned with the actual moral preferences of stakeholders.

The necessary precondition for moral prospection to commence is that organizational leaders must be aware of the moral content of a given decision, a decision-context referred to here as moral quandary. Organizational reputation concerns serve as the primary motivation to understand how the organization's moral decisions will be perceived by stakeholders, but the

process can also be motivated by a genuine desire to maintain one's moral character (Cohen & Morse, 2014; Goodwin et al., 2014; Helzer et al., 2014).

***Proposition 1:** When confronted with a moral quandary, organizational leaders will engage in a cognitive, prospective simulation of stakeholders' potential reactions to the organization's possible moral behaviors.*

Evidence supporting Proposition 1 can be found across research in management and psychology on the role of impression management concerns in the moral domain. Observers infer traits such as trustworthiness, warmth, and competence from the decisions people make in explicit moral dilemmas (Everett et al., 2016; Rom et al., 2017), and the people making such decisions are aware of these attributions and will strategically adjust their behavior accordingly (Rom & Conway, 2018). While psychological research in this domain is limited, it echoes multiple theorists who argue that moral norms and behaviors are fundamentally social phenomena meant to facilitate social bonds and communicate social information (Fiske & Rai, 2015; Haidt, 2012). Within organizations, impression management concerns often drive citizenship behaviors (Bolino, 1999), and employees will direct such behaviors at their supervisors in order to receive better ratings (Bolino et al., 2006). Similarly, CSR engagement among organizational leaders is often driven by perceived pressure from stakeholders (Stevens et al., 2005; Wang et al., 2015), and meta-perceptive concerns regarding how consumers will perceive an organization's CSR initiatives affects how employees perceive such initiatives (Panagopoulos et al., 2016). Collectively, these findings provide evidence for Proposition 1 by

highlighting how considerations of moral behavior can be driven by concerns over how one will be judged morally by others.

When leaders begin the process of moral prospection they first intuitively and iteratively attempt to cognitively simulate how stakeholders would perceive and react to a particular moral decision. This cognitive moral simulation is dynamic, affective, fast, and future-oriented (Gilbert & Wilson, 2007; Seligman et al., 2013), involving the simulation of multiple stakeholders (Reynolds et al., 2012; Wang et al., 2015), and how they may react to different decisions on the part of the prospector. Critically, the moral simulation involves forecasting how stakeholders will perceive the *organization*, not the leader engaging in the simulation. This moral simulation involves imagining the mental states (i.e. thoughts, feelings, attributions) of specific stakeholders, and how those mental states may drive stakeholder reactions to organizational behavior. In simulating those mental states, prospectors will be subject to the corporate personhood bias, where they fail to account for stakeholders' only partial anthropomorphization of organizations.

Proposition 2: *When simulating the mental state inferences that stakeholders will make toward their organization, leaders will be subject to the corporate personhood bias, where they overestimate the level of anthropomorphization, specifically experiential mind perception, stakeholders engage in towards the prospector's organization.*

Evidence for Proposition 2 can be found in research demonstrating that people hold corporations to different moral standards than individuals (Haran, 2013; Jago & Laurin, 2017; Jago & Pfeffer, 2018; Plitt et al., 2015), and that organizations and groups are perceived as less

capable of emotions than individuals (Aaker et al., 2010; Cooley et al., 2017; Kervyn et al., 2012; Knobe & Prinz, 2008). Additionally, Rai & Diermeier (2015) provide direct evidence for the existence of the corporate personhood bias. They demonstrate that while the general population perceives corporations as low in experiential mental states and high in agentic mental states, senior executives perceive hypothetical organizations as having equal levels of experiential and agentic mental states as human targets. While Rai & Diermeier (2015) found that senior executives were somewhat aware of the discontinuity between their judgements and the judgments of the general public, the context was that of perceptions towards hypothetical companies in experimental vignettes, not the companies for which these senior executives belonged. The theory of moral prospection argues that the corporate personhood bias arises from identification with one's organization, and as such the experimental measures from Rai & Diermeier (2015) do not directly reflect the context outlined by the theory. Furthermore, the established relationship between organizational identification and positive evaluations of CSR initiatives (De Roeck & Maon, 2016; Farooq et al., 2017; Glavas & Godwin, 2013), and the inaccuracy of meta-perceptive judgments regarding how individuals believe others perceive them (Chambers et al., 2008; Pronin, 2008; Vazire & Carlson, 2011) provides support for Proposition 2's assertion that moral prospectors will be subject to, and unaware of, the corporate personhood bias in their simulation of stakeholders' perceptions.

During the cognitive moral simulation prospectors will generate moral meta-perceptions: discrete estimations of how stakeholders might react towards a particular organizational decision. However, due to the corporate personhood bias many of these moral meta-perceptions will be inaccurate and overly positive. That is, moral meta-perceptions will consistently underestimate stakeholders' attributions that a decision is immoral and socially irresponsible. Specifically,

moral meta-perceptions will be inaccurate in the domain of perceived good intentions, and this misestimation will drive further inaccuracies in meta-judgments of perceived harm caused by the organization's behavior, perceived organizational culpability, and perceived victim-status of parties affected by the organization's behavior. This constellation of inaccurate, and overly positive, moral meta-perceptions will ultimately lead to an underestimation of the extent to which stakeholders will perceive a particular decision as immoral and socially irresponsible.

***Proposition 3a:** The corporate personhood bias in the cognitive moral simulation will cause an inaccurate overestimation of the extent to which stakeholders will perceive a particular organizational decision as motivated by altruistic and sympathetic intent.*

***Proposition 3b:** Inaccurate estimations of perceived positive intent will drive inaccurate, and overly positive, estimations of perceived harm caused, organizational culpability, and affected party non-complicity, in the domain of the organization's decision.*

***Proposition 3c:** Inaccurate moral meta-perceptions of attributed intent, harm caused, culpability, and victim-status will lead to inaccurate, and overly positive, meta-judgments of the extent to which a particular decision is perceived as cumulatively immoral and socially irresponsible by stakeholders.*

Propositions 3a, 3b and 3c are supported by research on attributions of CSIR, the role of mind perception in moral judgments, and the role of intent attributions in driving moral judgments. Proposition 3a is derived from the nature of the corporate personhood bias (Rai &

Diermeier, 2015), where moral prospectors fail to estimate the reduced level of experiential mind perception stakeholders engage in toward the organization. Given the relationship between agentic and experiential mind perception (Epley & Waytz, 2010; Gray et al., 2007), agents that are high in agency and low in experience are perceived as capable of having and acting on intentions, due to their high agency, but not capable of possessing intentions characterized by emotion, altruism, and empathy, due to their low experience (Kervyn et al., 2012; Knobe & Prinz, 2008; Rai & Diermeier, 2015; Schein & Gray, 2018). As such, the corporate personhood bias will drive over-estimation of stakeholder attributions of positive (i.e. sympathetic, altruistic, selfless) intention behind an organization's decision.

Proposition 3b is supported by evidence that judgments of intent, when observing another's moral behavior, affect judgments of harm caused (Ames & Fiske, 2015; Chakroff et al., 2016; Gray, 2012), judgments of organizational culpability, which in Lange & Washburn's (2012) model encompasses perceived causality (Cushman, 2015; Pizarro et al., 2003) and blame (Cushman, 2008; Schein & Gray, 2014; Young & Saxe, 2011), and judgments of victimhood-status (Gray & Wegner, 2009; Niemi & Young, 2014, 2016). As such, inaccuracies in meta-perceptions of perceived intent (Proposition 3a) will drive underestimation of perceived harm caused by the organization's decision, the perceived causal connection between the organization's decision and the outcomes engendered by the decision, the perceived organizational blameworthiness related to those outcomes, and the perceived non-complicity of parties affected by the organization's decision. Proposition 3c is supported by evidence demonstrating that judgments of intent, harm, culpability, and victim-status are all critical components of final judgments as to whether an organization's behavior is immoral and socially irresponsible (Cushman, 2008, 2013; Gray et al., 2014; Gray & Wegner, 2009; Jago & Laurin,

2017; Jago & Pfeffer, 2018; Lange & Washburn, 2012; Vlachos et al., 2013; Young et al., 2007). As such, if prospectors have inaccurate, and overly positive, meta-perceptions in the domains of intent, harm, culpability, and victim-status, they will generate similarly inaccurate, and overly positive, cumulative meta-judgments of perceived immorality and social irresponsibly.

Once moral prospectors have generated inaccurate moral meta-perceptions toward a potential moral decision on the part of the organization, the moral meta-perceptions will engender a process of moral learning, which leads to a state of moral overconfidence. Moral overconfidence is defined as over-precision (i.e. undue confidence in the accuracy of a judgment, see: Moore, Tenney, & Haran, 2010) in one's moral meta-perceptions, and the updating of one's own moral preferences to be congruent with one's moral meta-perceptions. In an effort to attenuate the dissonance surrounding their moral preferences related to the moral quandary at hand, prospectors update their personal moral preferences to reflect the (inaccurate) meta-perceptions they have generated. This process is fundamentally one of *learning*. Prospectors do not simply update their beliefs about stakeholders' attitudes, they change their personally held beliefs about the nature of the moral quandary as a function of the simulated stakeholder feedback. A specific decision which, pre-prospection, may have been perceived as morally ambiguous will, due to moral overconfidence, be considered congruent with one's moral preferences and therefore moral once prospection is complete. Moral overconfidence arms the prospector with newly formed moral preferences which seemingly provide clarity in respect to the moral quandary. Moral prospectors will then make decisions aligned with these preferences, and such decisions will engender unanticipated moral backlash from stakeholders.

Proposition 4: *Inaccurate moral meta-perceptions will foster a state of moral overconfidence, where over-precision in the accuracy of one's moral meta-perceptions leads prospectors to update and codify their own moral preferences surrounding the decision at hand.*

Proposition 5: *Moral overconfidence will lead prospectors to make a moral decision aligned with their updated moral preferences, and this decision will engender moral backlash from stakeholders.*

Support for Proposition 4 can be found in the burgeoning literature on the role of prospection in moral learning. Prospection's role in moral learning is intuitive and empathetic (Railton, 2017), giving rise to empathy-driven systems of reinforcement learning (Cushman, 2013). Prospection can foster anticipatory guilt (Ahn et al., 2014; Baumeister et al., 1994), and future-oriented episodic simulations can foster pro-sociality, and function similarly to past-oriented episodic memory in fostering pro-social behavior (Benoit & Schacter, 2015; Gaesser & Schacter, 2014; Brendan Gaesser et al., 2017). Similarly, the temporality of moral behavior (i.e. past vs. future behavior) affects moral judgments (Caruso, 2010), as does the level of visual imagery engaged in when considering a moral dilemma (Amit & Greene, 2012). Nonetheless the forecasting accuracy of intuitive and empathic prospective judgment has been questioned by scholars (Gilbert & Wilson, 2007; Greene, 2017), providing further evidence for the inaccuracies detailed by the theory of moral prospection. While the construct of moral overconfidence has yet to be directly investigated by scholars, indirect evidence across the literature on moral learning

provides support for the link between moral prospection and the development of moral preferences proposed by Proposition 4.

Evidence for Proposition 5 exists across CSR and marketing scholarship, along with research linking moral preferences with moral behavior. A large body of psychological scholarship examines the relationship between moral beliefs/preferences and moral behaviors (see for review: Bartels, Bauman, Cushman, Pizarro, & McGraw, 2014; Haidt & Kesebir, 2010; Treviño, den Nieuwenboer, & Kish-Gephart, 2014). People have a generalized preference for engaging in moral behavior (Capraro & Rand, 2018), although moral principles are often invoked for rationalization purposes (Uhlmann et al., 2009). The motivation to maintain a moral identity can also drive moral behaviors (Aquino & Reed, 2002; Reynolds & Ceranic, 2007; Strohmingner & Nichols, 2014), as managers who self-report stronger moral values are more likely to engage in socially responsible behavior (Crilly et al., 2008). In regard to moral backlash to organizational behavior, there are high levels of incongruence between investors and stakeholders (Groening & Kanuri, 2013), and between consumers and organizations (Öberseder et al., 2013; Skarmeas & Leonidou, 2013) in how CSR initiatives and outcomes are perceived. Among consumers these incongruences cause strong moral emotions (Grappi et al., 2013) which lead to consumers' willingness to punish brands for perceived irresponsibility (Palazzo & Basu, 2007; Puzakova et al., 2013; Sweetin et al., 2013). Consumers will also express outrage at immaterial harms such as high corporate salaries (Tannenbaum et al., 2011).

Proposition 5, and the role of moral overconfidence in codifying moral preferences, helps explain why CEOs are often unapologetic in their initial reactions to scandals. While there is little systematic empirical research on the effectiveness of CEO apologies (Koehn & Goranova, 2016), the case of Starbucks' Race Together campaign provides a clear example of moral

overconfidence post-scandal. Then CEO Howard Schultz, in his letter announcing the end of the Race Together campaign only a week after it began, included no apologies or indications of regret, instead proclaiming that “We leaned in because we believed that starting this dialogue is what matters most. We are learning a lot. And will always aim high in our efforts to make a difference on the issues that matter most” (Starbucks Corporation, 2015b). It was later revealed that Starbucks did not perform any market research in the leadup to the campaign, and in an interview three months after the campaign’s end Schultz said of the campaign “We made a tactical error. So what? We’re moving forward” (Carr, 2015). While we cannot know Schultz’s true feelings regarding the failed Race Together campaign, it is difficult to construe this sentiment that the campaign was merely a “tactical error” as one of genuine contrition or acceptance that most perceived the Race Together campaign as socially irresponsible (Logan, 2016). This behavior is congruent with the theory’s conception of how moral overconfidence both codifies one’s own moral preferences and prevents prospectors from updating their moral meta-perceptions.

The dynamics detailed by Propositions 1 through 5 constitute the theory of moral prospection. While many of the propositions are indirectly supported by research across moral psychology and CSR, the novel nature of the phenomenon means that many components of the theory have not been the subject of direct empirical inquiry. Nonetheless, I argue that the overall phenomenon is intuitive to understand. Just as people attempt, sometimes accurately and other times inaccurately, to understand whether other people perceive them positively or negatively in non-moral domains (Vazire & Carlson, 2010), leaders in organizations face the same meta-perceptive perils. Broadly, this process of attempting to understand how stakeholders perceive one’s organization encompasses a range of reputational concerns (Lange et al., 2011). Here I

focus more specifically on the moral domain and attempt to elucidate the cognitive mechanisms underlying why leaders within an organization may fail to predict how their behavior will be perceived morally.

Discussion

The theory of moral prospection makes three important contributions to management scholarship. First, it highlights a novel phenomenon that contributes to real-world corporate social irresponsibility and scandal. Specifically, the theory presents insights into what occurs when someone has overcome moral blindness. Scholars in behavioral ethics have uncovered the myriad ways in which our cognition can blind us to the moral nature of our judgments and behaviors in organizations, including moral fading (Tenbrunsel & Messick, 2004), moral disengagement (Moore, 2008, 2015), and moral licensing (Blanken et al., 2015; Klotz & Bolino, 2013), along with a host of organizational and social factors that contribute to unethical behavior (Ashforth & Anand, 2003; Lees & Gino, 2017; Moore & Gino, 2013). Nonetheless this literature has yet to investigate what occurs when people overcome their “bounded ethicality” and still are not certain as to their own moral compass. The theory of moral prospection provides a novel avenue for behavioral ethics scholars to understand how cognitive biases may drive immoral behavior even when the decision-maker is well aware of the moral content of the decision and is motivated to act morally. The theory of moral prospection also builds upon a burgeoning literature in moral psychology demonstrating the role meta-beliefs play in shaping moral preferences and behaviors (Railton, 2016; Rom et al., 2017; Rom & Conway, 2018), and highlights the way in which these processes may be biased, leading to the learning of incongruent moral norms. Moral psychology has often studied moral judgments and moral

behaviors separately (Teper et al., 2011), and the theory of moral prospection provides a framework for understanding the direct connection between the two.

The theory of moral prospection also contributes to scholarship on corporate social irresponsibility. While there is a large literature on the psychological micro-foundations of CSR, the vast majority of it has focused on the antecedents of and reactions to socially responsible behavior, and scholars in this field have called upon researchers to better understand the roots of corporate social *irresponsibility* (Gond et al., 2017; Murphy & Schlegelmilch, 2013). Some work in this domain has found that leaders' desire for power can contribute to CSIR, but that this can be mitigated by shared organizational leadership (Fabrizi et al., 2014; Pearce & Manz, 2011). The theory of moral prospection builds on this research by providing a cognitively-based model of how leaders come to engage in behavior which evokes moral backlash. The role of meta-perceptions too has begun to attract the focus of CSR scholars. Panagopoulos, Rapp, & Vlachos (2016) find that employees' meta-beliefs regarding how they think the firm's CSR efforts are perceived influence their own perceptions of those efforts. The theory of moral prospection provides CSR scholars with an avenue for future research on the micro-foundations of CSIR and highlights the role meta-perceptions regarding one's own organization play in shaping irresponsible behavior. CSR research also highlights the role of leaders' beliefs and values in driving CSR engagement (Crilly et al., 2008; Hibbert & Cunliffe, 2015), and moral prospection's focus on moral learning provides one explanation as to how organizational leader may develop these moral beliefs.

Moral prospection's second major contribution to management scholarship is the integration of moral psychological research on anthropomorphization with CSIR research on how organizations are perceived. Marketing scholars have highlighted the role of

anthropomorphization in brand perceptions (Aggarwal & McGill, 2012; Puzakova et al., 2013), and psychologists have studied how these processes affect perceptions toward individuals and groups (Gray et al., 2012; Waytz & Young, 2012), yet research examining how anthropomorphization affects moral perceptions of organizations is limited to a handful of recent studies (e.g. Jago & Laurin, 2017; Jago & Pfeffer, 2018; Rai & Diermeier, 2015), and there have been no attempts to integrate this research with CSIR scholarship. Here we integrated Lange & Washburn's (2012) theoretical framework for attributions of CSIR with dyadic morality theory's framework outlining the cognitive mechanisms driving moral judgments (Schein & Gray, 2018), along with research on the role of perceived intent in driving such moral attributions (Ames & Fiske, 2015; Chakroff et al., 2016). As such the theory of moral prospection provides the foundation for researchers to understand moral perceptions toward organizations in a way that is both sensitive to the complexities of moral cognition while also specific to the nature of organizational perceptions and behaviors, rather than just moral perceptions toward individuals. The anthropomorphization framework, with its distinction between agentic and experiential mind perception, may also help disentangle conflicting results surrounding why organizations seem to be held to higher moral standards than people (e.g. Jago & Pfeffer, 2018), but are also able to engage in behavior which is considered acceptable for corporations but immoral for individuals (e.g. Haran, 2013).

An important caveat to the theory of moral prospection is that it draws on findings related to the anthropomorphic perceptions of for-profit companies. While it has yet to be directly tested, research suggests that anthropomorphization toward non-profit organizations may differ substantially from how for-profit companies are perceived (Aaker et al., 2010; Kervyn et al., 2012). Nonetheless, the theory of moral prospection can accommodate this difference between

non-profits and for-profits by reconsidering how the corporate personhood bias manifests within non-profit organizations. In for-profit organizations the corporate personhood bias is defined as an overestimation of the extent to which outsiders perceive the organization as capable of emotion and feeling (experiential mind perception). However, Aaker et al. (2010) suggests that non-profits are perceived as high in experience but low in agency, the opposite cognitive template of for-profit organizations. As such, in a non-profit organization the corporate personhood bias may manifest as an overestimation of perceived competency and control (agentic mind perception). Given this possibility, future research should unpack how non-profit entities (e.g. governments, charities, hospitals, universities, etc.) are anthropomorphized, and how moral prospection and the corporate personhood bias may lead organizational leaders in non-profit settings to generate inaccurate moral meta-perceptions.

The theory of moral prospection's third contribution to management scholarship is that it does not simply highlight a yet studied phenomenon, it details the specific cognitive bias which leads to inaccurate moral meta-perceptions and ultimately immoral behaviors. The identification of the corporate personhood bias, and its role in fostering decisions that engender moral backlash from stakeholders, provides a clear avenue for scholars attempting to reduce CSIR via psychological and policy interventions (Epley & Tannenbaum, 2017; Zhang et al., 2014). It is difficult to say how easy it will be to "de-bias" the process of moral prospection. On the one hand, research on perspective-taking suggests that people are indeed able to simulate the mental states of others and consider situations that are counterfactual to their own perceptions (Galinsky & Moskowitz, 2000; Ku et al., 2015), suggesting that overcoming the corporate personhood bias may be as simple as a targeted perspective-taking intervention. However, perspective-taking is often flawed and inaccurate (Eyal et al., 2018), and asking someone inside an organization to

overcome the corporate personhood bias is essentially asking them to dehumanize their own organization. Asking leaders in organizations to see their own organization as less human may be futile, as strong identification with the organization may prevent leaders from dehumanizing a group for which they identify with and value. And even if leaders can successfully take the perspective of those outside the organization as such, the intuitive nature of moral prospection and the corporate personhood bias mean that such effortful considerations may not affect the automatic processing involved in generating moral meta-perceptions. Additionally, the theory of moral prospection argues that the process ultimately leads to moral overconfidence and the codification of the belief that a decision is indeed moral. This suggests that the most successful interventions at de-biasing moral prospection would occur *before* the prospector acquires the inaccurate meta-perceptions which lead to moral overconfidence. This possibility is one of critical importance for future research.

While the theory of moral prospection strives for parsimony in its explanation of the cognitive processes underlying CSIR, several questions surrounding how moral prospection functions in the real-world of organizations remain. For example, how moral prospection unfolds when considering the multiple, and often conflicting, preferences of different stakeholders (Rayton et al., 2015; Reynolds et al., 2012; Turker, 2009; Wang et al., 2015) is a question not directly considered by the theory. Leaders engaging in moral prospection may privilege certain stakeholders over others, leading to variance in the extent to which their meta-perceptions are accurate or applicable to certain stakeholders (Groening & Kanuri, 2013). Additionally, if the corporate personhood bias arises from identification with the organization, it is likely that leaders will be more accurate when the stakeholder they are simulating also has relatively high levels of identification with the organization, such as employees. As such, the theory would predict that

leaders' meta-perceptions regarding employees and investors will be relatively more accurate than their meta-perceptions regarding the general public, customers, and regulatory and governing bodies. The theory would also predict that leaders who do not strongly identify with their organization will have more accurate meta-perceptions and will therefore be less likely to engage in CSIR. This possibility, and the dynamics surrounding moral prospection and stakeholder management, are fruitful avenues for future research.

While the theory of moral prospection focuses on the cognitive processes underlying the decisions of individual leaders, many moral decisions in organizations are made by groups of people, such as corporate boards. As such, group processes are likely to affect individuals' moral prospection in organizational contexts. There is surprisingly little research on how team dynamics affect moral decision making (but see: O'Leary & Pangemanan, 2007; Yang, Ji, & O'Leary, 2017), and research on moral cognition has yet to systematically investigate the role of social and organizational forces in moral judgments (Lees & Gino, 2017), although the topic has received more attention in the behavioral ethics literature (see: Moore & Gino, 2013; Treviño et al., 2014). Additionally, people are unlikely to average multiple meta-cognitive judgments (Fraundorf & Benjamin, 2014), suggesting that moral overconfidence may prevent prospectors from updating their meta-perceptive beliefs when others in the organization provide new meta-information. How team and organizational level processes affect moral prospection, especially in the context of multiple people engaging in the same prospective process simultaneously (such as in a team), is an important area of research for scholars to pursue.

A final consideration regarding the theory of moral prospection relates to the proportion of corporate scandals that are the result of moral prospection. This, of course, can only be answered through empirical investigation. I am not arguing that moral prospection is the cause of

all, most, or even a majority of CSIR and scandal. Many explanations for corporate misbehavior existing already, including economic self-interest (Carson, 2003), structural incentives (Coffee, 2005), and bounded ethicality (Bazerman & Gino, 2012), just to name a few. Nonetheless, I argue that there are many instances where organizations clearly failed to predict that their behavior would be perceived as unethical or irresponsible, or at the very least failed to predict the level of moral backlash. Ultimately, CSIR is the results of multiple processes, of which moral prospection is a contributing factor, and one that has yet to receive the attention of organizational and psychological scholars.

Conclusion

The theory of moral prospection highlights a novel and yet studied psychological phenomenon which contributes to corporate social irresponsibility and moral backlash from stakeholders. The theory utilizes insights from across moral psychology and CSIR research to elucidate the cognitive processes underlying why organizational leaders fail to predict how their organization will be perceived morally, and details how those inaccurate meta-beliefs lead to immoral decisions. It highlights the domains in which moral meta-perceptions will be inaccurate and provides an avenue for future psychological interventions by identifying the corporate personhood bias in prospectors' judgments: all with the ultimate goal of inspiring research into organizational scandals and understanding how they can be prevented.

Chapter 3

Do managers know how their decisions are perceived? Measuring the accuracy of CSR-related group meta-perception

Jeffrey Lees

Abstract

Drawing from samples of managers with corporate social responsibility (CSR) and reputation management experience, we assess the accuracy of managers' group meta-perceptive judgments in CSR-related decisions domains. Using a novel written recall task where managers are asked to write about past decisions related to CSR, diversity, and/or institutional values, we directly access managers meta-perceptions regarding the motives that the average individual would attribute to those decisions and how it might affect perceptions of the company. Managers' written accounts are then provided to a nationally representative sample of Americans who judge the managers' decisions, allowing for a direct test of managerial meta-perceptive accuracy. We adopt a componential approach to understanding judgment accuracy, allowing us to disentangle multiple judgment components that may be driving (in)accuracy in managers' group meta-perceptions. Our findings will provide a detailed descriptive account of how, and in what domains, managerial group meta-perceptions are inaccurate, and what managerial traits predict greater or less accuracy. This research will lay the groundwork for future targeted interventions designed to improve managers' decision-making and judgment accuracy.

Introduction

Judgments of how others perceive one's group – group meta-perception – are an essential aspect of leaders' decision-making processes surrounding group-level behaviors. Nonetheless, the possibility of inaccuracy and bias in group meta-perceptive judgments (Carlson et al., 2011; Vazire & Carlson, 2010) represents a challenge to leaders hoping to foster and maintain cooperative organizational relationships and a positive organizational reputation. As issues surrounding corporate social responsibility (CSR) become more salient to organizations and the public (Sweetin et al., 2013), organizational leaders are confronted with the need to better understand the normative attributions relevant stakeholders will make toward organizational decisions (Lange & Washburn, 2012; Vlachos et al., 2010). As such, the central question this research seeks to answer is to what extent are organizational leaders (in)accurate and (un)biased in group meta-perceptive judgments related to organizational social responsibility, what factors moderate accuracy, and what consequences might arise from inaccuracy.

This research is grounded in the experience of real-world organizations by directly accessing a sample of managers and eliciting from them decisions they have faced surrounding CSR within their organization. Group meta-perceptive judgments from these managers are then compared to a nationally representative sample, where participants provide their perceptions and attributions toward the decisions described by the managers. This design allows for a direct test of group meta-perceptive accuracy and bias, along with the ability to test for moderators of and outcomes associated with accuracy. Rather than relying on highly stylized and fictional scenarios that are commonly used, and increasingly criticized (Graham, 2014; Hofmann et al., 2014) in social psychology, we employ a grounded approach where we directly solicit written experiences

from organizational leaders regarding decisions they have confronted related to social responsibility.

This research utilizes recent methodological advances in the study of meta-perception and accuracy in social judgment. Past work has been criticized for reliance on problematic analyses of raw difference-scores (Barranti et al., 2017; Shanock et al., 2010), the theoretical and empirical conflation of accuracy and bias (West & Kenny, 2011), and failure to control for stereotypic judgments in the assessment of target-specific accuracy (Biesanz, 2010; D. Wood & Furr, 2016). In our examination of accuracy, bias, and moderators thereof, this research utilizes the social accuracy model of interpersonal perception (Biesanz, 2010), which relies on hierarchical linear mixed-effects modeling to disentangle accuracy and bias, while controlling for stereotypic judgments, all in a single framework that provides high levels of statistical power and avoids issues of multiple statistical comparisons.

This research is conducted across two phases. In phase one the investigators gather a sample of organizational leaders who consent to sharing instances where they confronted an organizational decision related to socially responsible (or irresponsible) corporate behaviors. Importantly, these leaders agree to (1) share their story with us, and (2) allow this story to be used as stimuli in phase two, where it is presented to survey participants in a nationally representative sample along with attributional items on the same dimensions which leaders provided their meta-perceptions in phase 1. The data collected in phase one and phase two will then be merged into a single dataset and analyzed using the statistical techniques developed by the social accuracy model.

Methods

Overview: Measuring Accuracy

Both phases 1 and 2 will utilize the social accuracy model of interpersonal perception (SAM, Biesanz, 2010) to analyze accuracy. The SAM is an extension of the classic social relations model (SRM. Kenny & Albright, 1987; Kenny & DePaulo, 1993) and utilizes mixed-effects modeling to examine interpersonal judgment and meta-perceptive accuracy. Whereas the SRM requires a fully crossed round-robin experimental design, the SAM is robust to unbalanced designs, for example when there are fewer actors than observers in a dataset. Moreover, the SAM allows for the analysis of individual-level moderators of accuracy (e.g., managers' years of experience) while still providing the decomposition of explained-variance across actors and observers that the SRM provides.

A critically important aspect of the SAM framework is the ability to examine multiple components of accuracy. Accuracy is often analyzed using raw difference scores, but this approach is statistically problematic and collapses across multiple components of accuracy (Barranti et al., 2017; Edwards & Parry, 1993). One such component of accuracy is mean-level accuracy: *within* a given judgment, is the judgment equal to, below, or above the true value. A second component of accuracy is rank-order accuracy: *across* multiple judgments, does the judge get the rank-order relatively right. In other words, is there a linear relationship between judgments and the true-values, sometimes referred to as profile-agreement. There's also normative-accuracy (sometimes called stereotypic accuracy, Furr, 2008): does the judge know the true-population mean related to the judgment at hand, and are they using this knowledge to guide judgments of specific-others. For example, a CEO at a bank attempting to judge their company's reputation may not have much insight into their *specific* company's reputation (low "distinctive" accuracy), but they may have an accurate understanding of how the *average* Bank is perceived (high normative accuracy), and they can use this knowledge to anchor their

judgments of their specific company's reputation. Normative-accuracy too has a linear component and mean-level component. Lastly, there are biasing forces that can pull on judgment and make it more or less accurate (West & Kenny, 2011), for example self-perception often biases meta-perception (Vazire & Carlson, 2011) and individuals frequently engage in projection of their own traits/preferences (Ames, 2004), and these biasing forces can both affect rank-order accuracy and mean-level accuracy. The SAM allows for an examination of *all* these components of judgment accuracy within a unified mixed-effects regression framework.

Phase 1

Phase 1 involves the collection of written accounts (“stories”) from a managerial sample regarding past decisions they have made related to social responsibility, diversity, and/or institutional values. Managers are asked about their motivations for this past behavior, and their appraisals of the behavior. They are also asked for accompanying meta-perceptions of how they think the average person would perceive their behavior. Data from phase 1 will be combined with data from phase 2 to allow for a direct test of managers’ meta-perceptive accuracy. *As of February 2020, data collection for phase 1 is ongoing.*

Participants: 30 managers with experience in reputation management and/or ESG/CSR investing will be recruited through the professional network associated with Orenda Solutions (<https://orendasolutions.com/>), a small tech start-up based in Toronto, Canada. Orenda uses big-data to help companies understand their reputation and the public’s sentiment toward the company, and is helping to recruit study participants from their network of clients and contacts in Canada and the United States. To incentivize managers to participate, managers are asked to list a charity of their choice they would like entered into a raffle for several \$300 donations. The

study is advertised as taking 5-10 minutes to complete, and the survey is facilitated through the Qualtrics survey platform.

Procedure: After providing informed consent participants are asked to describe, in writing, a past managerial decision they have made regarding social responsibility, diversity, and institutional values (see Appendix for the prompt's exact language). Participants are asked to write 3-5 sentences and provided an open text box. Participants are explicitly told that they will be asked, later in the survey, how they believe someone else might judge their behavior (i.e., meta-perceptions). They are also explicitly told not to include any information that could identify them or their organization in their story.

On the following page participants are asked for their meta-perceptions. On this page they are given the exact language of the story they previously wrote, allowing them to see their story as they judge how it might be perceived by others. Meta-perceptions are assessed across 13 unipolar motive items (e.g., selflessness, company interests, fairness), 3 bipolar moral judgment items related to the behavior (e.g., unethical – ethical), 3 unipolar items measuring sympathy toward the organization, and 5 unipolar items measuring attributions toward their organization (e.g., trustworthy, socially responsible). On the following page participants are asked for their self-perceptions toward their own behavior, specifically along the 13 motive and 3 moral judgment items above. All items use 7-point Likert scales throughout, and the order of items with meta- and self-perceptions are counterbalanced.

Participants are then explicitly asked if they consent to the use of their story as stimuli in future research, and are assured that their identity will be kept completely confidential. They are then asked to complete several trait measures: trait perspective-taking and empathic concern (Davis, 1983), trait Machiavellianism (Dahling et al., 2009), and propensity to engage in

unethical workplace behaviors (Chen & Tang, 2006). Participants are then asked to provide personal and identifying information about themselves, and are reminded that all this information is kept confidential and providing it is optional. They are asked for their name, email, current organization and position therein, and what organization/position they were in at the time of the story they provided. Participants are then asked for their age, gender, years of industry experience, the charity they wish to enter into the raffle, and lastly are provided an open text box to leave comments.

Planned Analyses: Stories will be first read carefully by the research team to guarantee that (a) they meet a basic and reasonable level of comprehensibility, and (b) they do not include any identifying information. Any scenarios that do not meet these criteria, or which the participant denied consent to share, will not be used as stimuli in phase 2 and therefore will go unanalyzed. Most planned analyses related to the stories require data from phase 2 (see phase 2's Planned Analyses below). For the data collected in phase 1, summary statistics, Spearman correlations, and reliability statistics (Cronbach's alpha, where appropriate) will be calculated and reported across all measures.

Phase 2

Phase 2 involves taking the stories collected in phase 1 and using them as stimuli for a separate general-population sample of "observers" to judge and rate. This design allows for a direct test of meta-perceptive accuracy by comparing the actual-perception of observers in phase 2 to the meta-perceptions of managers in phase 1.

Participants: 163 participants will be collected from a nationally representative sample using Qualtrics survey panels and demographic quotas matched to the US census along the following characteristics: age, gender, ethnicity, education, and income. This sample size was

determined based on guidelines from Barranti et al. (2017) regarding the sample needed to detect small effect sizes of judgment accuracy with 90% statistical power. The study will be advertised as taking ~20 minutes and participants will be paid a predetermined amount of credits through Qualtrics' internal credit system.

Procedure: After providing informed consent participants will read a randomly-selected written account of a past behavior of a “high-level manager in a for-profit business” (i.e., a story from phase 1; see Appendix for the prompt). They will be explicitly told that these accounts are true and were written by participants as part of this study. On the page displaying the story participants will be asked for their perceptions along the same exact items measured as meta-perceptions in phase 1: 13 unipolar motive items, 3 bipolar moral judgment items related to the behavior, 3 unipolar items measuring sympathy toward the organization, and 5 unipolar items measuring attributions toward the organization. The order of items will be counterbalanced. After rating the first story this process will repeat over a total of five randomly selected stories from phase 1. Afterwards participants will complete measures of trait perspective-taking and empathic concern (Davis, 1983), trait Machiavellianism (Dahling et al., 2009), and propensity to engage in unethical workplace behaviors (Chen & Tang, 2006). Participants will then answer basic demographic questions, be given the opportunity to provide open comment, and the survey will end.

Planned Analyses: Prior to merging the data from phase 1 with data from phase 2, summary statistics, semi-partial Spearman correlations (controlling for repeated-measures), and reliability statistics (Cronbach's alpha, where appropriate) will be calculated and reported across all measures. Phase 1 and 2 data will then be merged, with the meta-perception ratings for managers matched to the corresponding observer ratings toward the respective managers' stories.

Meta-perceptions will be true-mean centered within item, as suggested by Biesanz (2010). True-mean centering allows for interpretable, within item, intercept estimates. Any intercept estimate that significantly varies from zero can be interpreted as mean-level directional bias for that measurement.

Following the SAM, linear mixed-effects modeling will be used to assess multiple components of meta-perceptive accuracy: baseline meta-accuracy, distinctive meta-accuracy, meta-insight, mean-level directional biases in accuracy, and trait moderators of accuracy. Across all models we will adopt a maximal random structure approach (Barr et al., 2013), where all linear predictors will be modeled as random slopes within random intercepts for managers' stories.

Baseline Meta-Accuracy Model: In all models of meta-accuracy managers' meta-perceptions, across all judgments, are modeled as the dependent variable. In the baseline meta-accuracy model managers' meta-perceptions are regressed onto the actual-perceptions of observers, with managers' stories modeled as random intercepts with random slopes for observers' actual-perceptions. A significant positive slope for the meta- and actual-perception relationship would constitute evidence for baseline meta-accuracy.

Distinctive Meta-Accuracy Model: While valuable, the estimate observed in the baseline model can potentially provide an inflated estimate of accuracy (Wood & Furr, 2016), as it does not take into account known confounders of interpersonal judgment accuracy. One such confounder is the normative profile: the true population-level mean distribution of the judgments at hand. Here, the normative profile is the distribution of mean self-perceptions among managers (i.e., their actual motives for engaging in the behavior in their story). Across judgment contexts individuals tend to rely heavily on the normative profile (Biesanz & Human, 2010; Carlson,

2016; Furr, 2008), and in this context reliance on the normative profile represents managers using a sense of how the *average* company/manager is perceived by the general public to guide their meta-perceptions of how they specifically are perceived. As such, the distinctive meta-accuracy model will add the normative profile into the baseline meta-accuracy model as a second independent variable. By controlling for the normative profile, the relationship between meta- and actual-perceptions can be interpreted as distinctive meta-accuracy: to what extent are managers able to predict how they *specifically* are perceived. Additionally, the relationship between meta-perceptions and the normative profile will be interpreted as the extent to which managers rely on normativity when generating meta-perceptions, i.e. normative accuracy.

Meta-Insight Model: Another confounder of meta-perceptive accuracy is the transparency of managers' motives and self-perceptions. Conceptually, a meta-perceiver may be accurate not because they have "insight" (Carlson et al., 2011) into how they are actually perceived, but simply because their meta-perceptions track closely with their true motives/feelings ("transparency bias," Gilovich et al., 1998) and their true motives/feelings are relatively transparent to observers. As such, the meta-insight model will add manager's individual self-perceptions as a second independent variable. By controlling for managers' self-perceptions, the relationship between meta- and actual-perceptions can be interpreted as meta-insight: to what extent can managers accurately predict how they will be *misperceived* by observers. Additionally, the relationship between meta-perceptions and self-perceptions will be interpreted as the extent to which managers exhibit a transparency bias in judging the ability of observers to accurately perceive their motives/feelings.

Directional Bias Model: To examine mean-level directional biases within judgment we will add a categorical variable denoting the specific judgment-item into the distinctive meta-

accuracy model and interact it with observers' actual-perceptions in predicting meta-perceptions, then compute least-squared means for intercept estimates of every item. Because meta-perceptions are true-mean centered, any intercept estimates that significantly deviates from zero can be interpreted as directional bias, meaning that across all managers there is systematic over- or under-estimation of a specific judgment, relative to the true value. This analysis also allows for an examination of descriptive patterns of directional bias, such as whether over- and/or under-estimation seems driven by the valence of the items.

Trait Moderation of Accuracy Models: To examine trait moderators of meta-accuracy we will separately enter each manager's trait measure into the distinctive meta-accuracy model. For each separate model the trait measure will be entered as an interaction with both actual-perceptions and the normative-profile in predicting meta-perceptions. A significant interaction with actual-perceptions in predicting meta-perceptions will be interpreted as moderation of accuracy, meaning that managers higher/lower on the given trait are systematically more accurate in their distinctive meta-perceptive judgments.

Conclusion

While there is no doubt that managers must frequently consider how their decisions might affect the reputation of their company, little research has examined whether such forecasts are accurate, let alone how inaccuracy may drive reputationally damaging decisions. The research here will begin to provide scholars with a detailed descriptive account of the systematic ways in which such group meta-perceptive judgments may be directionally biased, inaccurate, and related to individual differences among managers. These findings will lay the groundwork for testable, confirmatory hypotheses regarding how inaccurate group meta-perceptions may drive poor decision-making on the part of managers by highlighting the specific domains in

which judgments are inaccurate. These findings will also begin to illuminate how behavioral interventions might increase group meta-perceptive accuracy among managers and contribute to better decision-making and managing of organizations' reputations.

Future Directions

The research presented herein represents some of the first empirical investigations into the accuracy of group meta-perception. While inspired by the long history of research in psychology on the accuracy of interpersonal judgments and meta-perception (Cronbach, 1955; Kenny & Albright, 1987; Laing et al., 1966), there are myriad empirical and theoretical questions remaining regarding group meta-perception as a phenomenon. Below I highlight several pressing questions raised by the work conducted as part of this dissertation.

A Theoretical Framework for Predicting Direction and Levels of Accuracy

This dissertation alone exemplifies how group meta-perceptive accuracy can vary wildly by social and judgment context. Chapter 1 argues that group meta-perception can be highly negative or relatively accurate depending on the nature of the group-level interaction, whereas Chapter 2 argues that self-anchoring biases can lead group meta-perception to be overly positive, and Chapter 3 argues that group meta-perception may have multiple components which can vary independently in severity and directionality. This diversity in judgment patterns parallels research on interpersonal accuracy which generally finds that accuracy is less a function of domain-general cognitive capacity and more a function of the social relations between the perceiver and the perceived (Carlson, 2016; Eyal et al., 2018; Zaki et al., 2008). Nonetheless, without a unified theoretical framework for predicting the conditions under which group meta-perceptions are inaccurate vs. accurate, overly positive vs. overly negative, or exhibit higher rank-order vs. directional precision (to name only a few relevant dimensions), it will be difficult to systematically understand how group meta-perceptions drive impactful decisions, or develop generalizable interventions designed to improve decision-making. Moving toward such a

unifying theoretical framework ought to be the primary goal of scholars researching group meta-perception over the coming years.

Developing Reliable Measures

Past work on meta-perception has been largely defined by two patterns: (1) a focus on interpersonal perception, and (2) a focus on trait/personality attributions. I would argue that part of the reason for this is the tremendous amount of work the field of personality has dedicated to validating their trait measures, such that the “Big 5” and HEXACO models of personality (Goldberg, 1992; Lee & Ashton, 2004) that dominate the meta-perception literature. This however presents issues for research on group meta-perception, namely the relative lack of well validated trait measures of group agents and the near complete lack of well validated *state* measures of things such as motive attributions or emotions one might make of a group agent. Even at the interpersonal level, meta-perceptions of *behaviors* are much less studied than meta-perceptions of stable traits. The use of only ad-hoc measures, as is the case throughout this dissertation, presents both an empirical challenge to research on group meta-perception, but more importantly will serve to prevent the accumulation of consistent and comparable knowledge needed to build toward more general theories of group meta-perception. This problem however is not unique to research on group meta-perception, as a such researchers interested in group meta-perception should actively collaborate with scholars studying group-perceptions/stereotypes, moral judgments, and behavioral attributions to develop and validate measures of both group traits and state-attributions.

How Does Individual and Group Meta-Perception Differ?

This research is directly inspired by and deeply indebted to the decades of theory, research, and methodological advances on interpersonal meta-perception. Said theory however

does not match perfectly onto group meta-perception as a cognitive phenomenon, which suggests that existing methods for measuring meta-perceptive accuracy might not perfectly capture the nature of group meta-perception. Individual meta-perception involves the cognitive simulation of two agents: one's self and a single specific other. But with group meta-perception there is a third agent: the group. How or whether the inclusion of this third agent may fundamentally alter the social cognition of meta-perception is not understood. For example, assumed similarity/projection is a robust and well documented phenomenon in social perception research (Ames, 2004; Cronbach, 1955; Hoch & Research, 1987; Robbins & Krueger, 2005). In an interpersonal context there is only one entity to project onto, but with group meta-perception there are two. Are group meta-perceivers projecting their own preferences onto the individual whose perceptions they are making forecasts of, or the group itself, or both? Are meta-perceivers disassociating themselves from the group in this three-way cognitive simulation? How might in-group biases effect group meta-perceptive forecasts, and are these in-group biases separate from preference-projection tendencies? Does one's relationship with their group, or standing therein, affect accuracy? It is not clear whether current methods such as the social accuracy model (Biesanz, 2010) and response surface analysis (Barranti et al., 2017) are best suited for answering these questions, and as such scholars pursuing research on group meta-perception will need to think critically about whether existing methods are sufficient to capture the full breadth of group meta-perception as a phenomenon.

Are There Differences Between Leaders and Group Followers?

Group leaders face meaningfully different reputational pressures than do the average group members, and as such there is reason to hypothesize that leaders' group meta-perceptions may differ systematically from those of the average group member. Chapter 2 herein suggests

that organizational leaders may be the most likely to exhibit a corporate personhood bias, which will in turn drive inaccurate and overly positive group meta-perceptions. However, given the evidence that social distance is a strong predictor of meta-perceptive inaccuracy at the interpersonal level (Carlson, 2016), one could also predict that leaders are on average more accurate in their group meta-perceptions than are the average group members. Understanding such differences is critical, as it might be the case that interventions designed to increase group meta-perceptive accuracy based on research using the average group member (e.g., the intervention presented in Chapter 1) may not generalize to group leaders.

Conclusion

My hope is that research on group meta-perception will help to build scholarly bridges between different literatures and disciplines. Each chapter of this dissertation is framed for a different audience, whether intergroup relations scholars, management scholars, or social perception scholars. Research on group meta-perception will be best served by integrating a diversity of disciplinary perspectives, theories, and methods. Going forward I urge all who might be interested in understanding group meta-perception to pursue that interest, bring to bear their unique perspectives, and contribute to the growing body of work on how we think others perceive the behavior of our group.

Supplementary Notes
Demographic Characteristics of Samples from Experiments/Studies 1-6

Demographic Characteristics: Experiment 1

Collected on Mechanical Turk

Total N = 408
 $M_{age} = 35.2$, $SD_{age} = 11.1$
 239 Women, 169 Men
 271 Democrats, 137 Republicans

	Actual-P Condition	Control Condition	Meta-P Condition		Female	Male
Democrat	95	82	94	Democrat	167	104
Republican	48	54	35	Republican	72	65

Demographic Characteristics: Experiment 2

Collected on Mechanical Turk

Total N = 286
 $M_{age} = 36.2$, $SD_{age} = 11.5$
 156 Women, 130 Men

	Actual-P Condition	Meta-P Condition
Female	87	69
Male	71	59

Demographic Characteristics: Experiment 3

Collected on Mechanical Turk

Total N = 499
 M_{age} = 35.1, SD_{age} = 11.9
 293 Women, 206 Men
 328 Democrats, 171 Republicans

	Actual-P Condition	Meta-P Condition		Female	Male
Democrat	165	163	Democrat	199	129
Republican	101	70	Republican	94	77

Demographic Characteristics: Experiment 4

Collected via Qualtrics Survey Panels

Experiment 4 was quota matched to census population characteristics such that the survey would be representative of the general American population. Below are the quotes utilized in data collection. We set out to collect N = 500, and Qualtrics purposefully oversampled to guarantee data quality. Total N = 536.

Quotas:

Gender:
 51% Female
 49% Male

21% \$100k - \$200k
 6% \$200k+

Age:
 32% 18-34
 34% 35-54
 34% 55+

Ethnicity:
 63% Non-Hispanic White
 12% Non-Hispanic Black
 17% Hispanic
 5% Asian
 3% American Indian/Alaskan Native/Other

Income:
 40% \$0 – \$50k
 33% \$50k – \$100k

Education:

41% HS Diploma/GED
 21% Some College (no degree)
 27% College Degree
 11% Graduate Degree

Political Affiliation:
 50% Democrat
 50% Republican

Below are the characteristics of the sample collected:

Total N = 536

Gender:

Female: 273 (50.9%)
 Male: 263 (49.1%)

Age:

18-34: 165 (30.8%)
 35-54: 189 (35.3%)
 55+: 182 (34%)

Income:

\$0 – \$50k: 213 (39.7%)
 \$50k – \$100k: 180 (33.6%)
 \$100k - \$200k: 109 (20.3%)
 \$200k+: 28 (5.2%)
 Prefer not to say: 6 (1.1%)

Ethnicity:

Non-Hispanic White: 344 (64.2%)
 Non-Hispanic Black: 61 (11.4%)
 Hispanic: 88 (16.4%)
 Asian: 26 (4.9%)
 American Indian/Alaskan Native/Other: 13 (2.4%)
 Prefer not to say: 4 (0.7%)

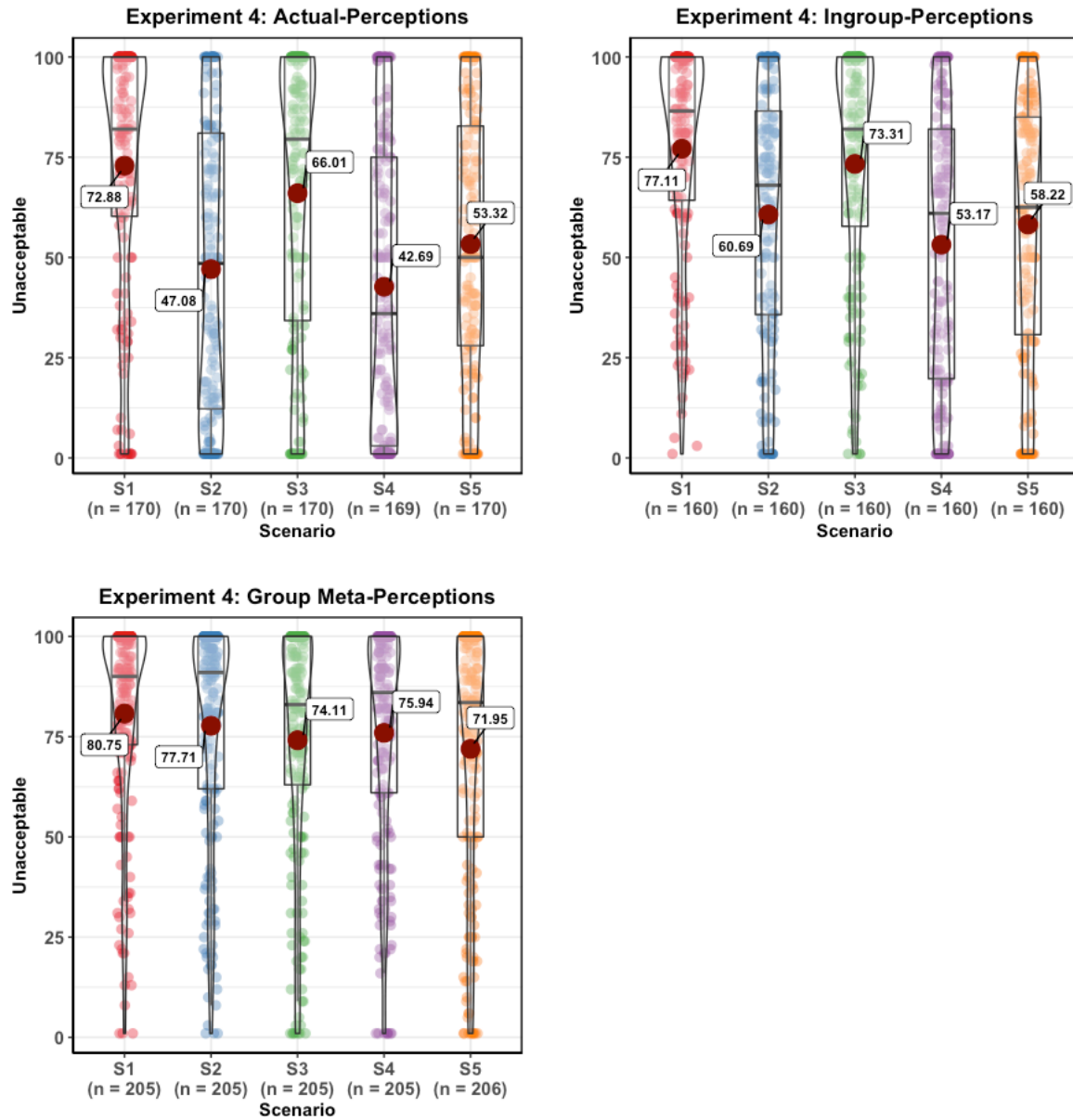
Education:

HS Diploma/GED: 194 (36.2%)
 Some College (no degree): 114 (21.3%)
 College Degree: 154 (28.7%)
 Graduate Degree: 72 (13.4%)
 Prefer not to say: 2 (0.4%)

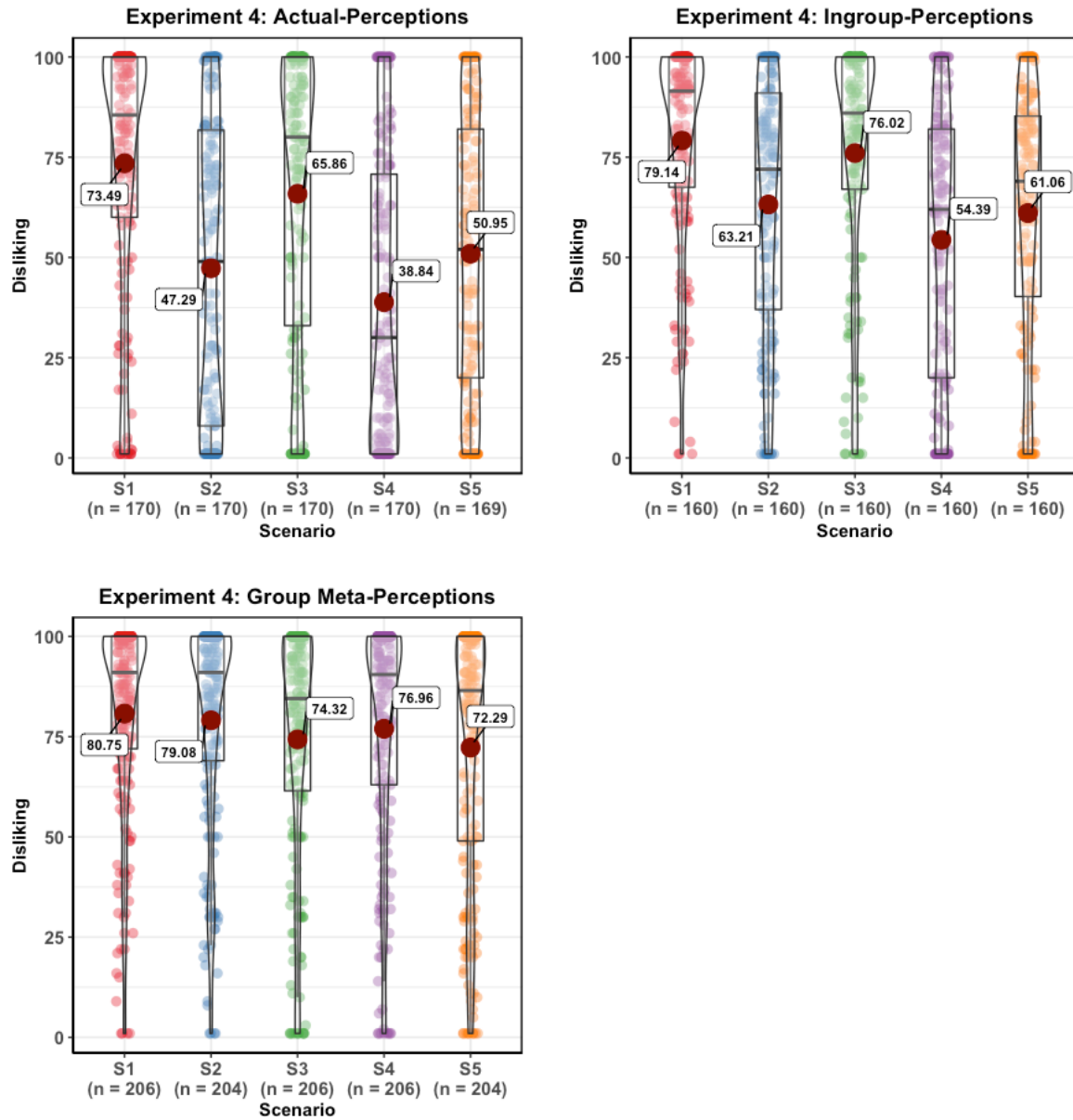
Political Affiliation:

Democrat: 269 (50.2%)
 Republican: 267 (49.8%)

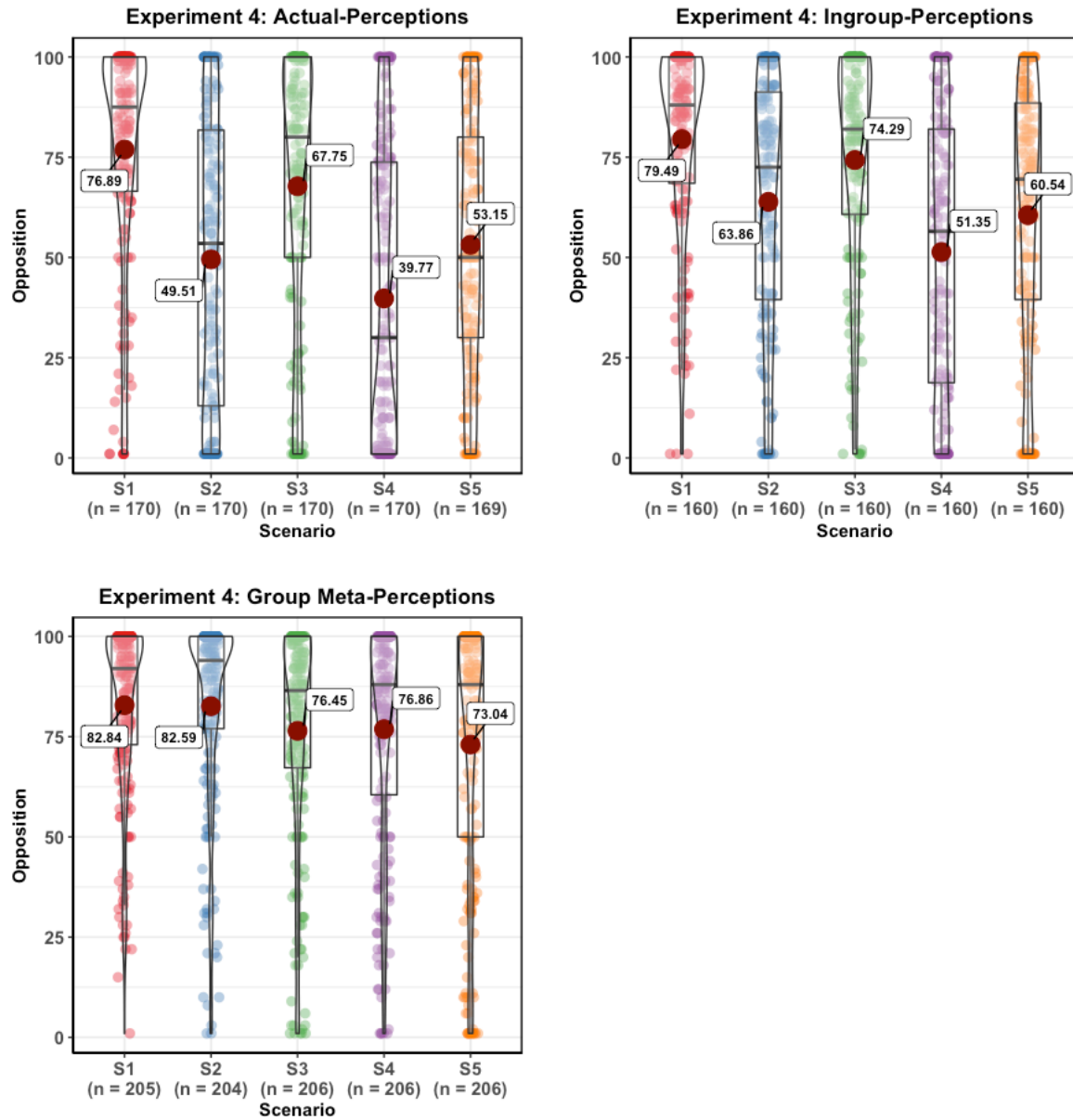
	Actual-P Condition	Ingroup-P Condition	Meta-P Condition		Female	Male
Democrat	82	79	108	Democrat	187	82
Republican	88	81	98	Republican	86	181



Supplementary Figure 1: Distributions of Unacceptability measure by condition and scenario in Experiment 4. Conditions (actual-perceptions, group meta-perceptions, and ingroup perceptions) are between-subjects, and within condition participants read and rated all five Scenarios (S1 – S5). Red dots and corresponding numbers are sample means, the boxplot center lines are sample medians.



Supplementary Figure 2: Distributions of Disliking measure by condition and scenario in Experiment 4. Conditions (actual-perceptions, group meta-perceptions, and ingroup perceptions) are between-subjects, and within condition participants read and rated all five Scenarios (S1 – S5). Red dots and corresponding numbers are sample means, the boxplot center lines are sample medians.



Supplementary Figure 3: Distributions of Opposition measure by condition and scenario in Experiment 4. Conditions (actual-perceptions, group meta-perceptions, and ingroup perceptions) are between-subjects, and within condition participants read and rated all five Scenarios (S1 – S5). Red dots and corresponding numbers are sample means, the boxplot center lines are sample medians.

Demographic Characteristics: Study 5

Collected on Mechanical Turk

Total N = 212

$M_{\text{age}} = 35.89$, $SD_{\text{age}} = 11.5$
 120 Women, 92 Men
 132 Democrats, 80 Republicans

	Female	Male
Democrat	80	52
Republican	40	40

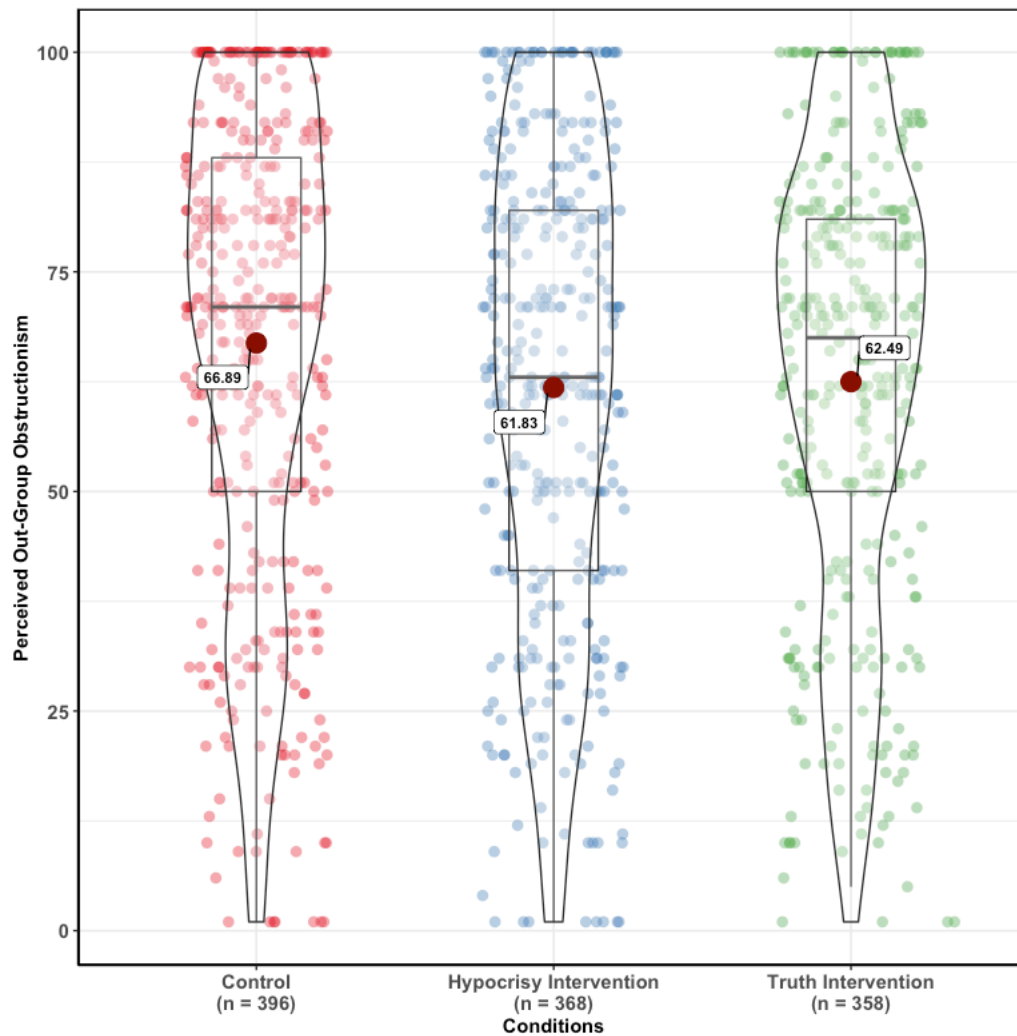
Demographic Characteristics: Experiment 6

Collected on Mechanical Turk

Total N = 1122
 $M_{\text{age}} = 35.1$, $SD_{\text{age}} = 11.6$
 642 Women, 480 Men
 704 Democrats, 418 Republicans

	Control	Hypocrisy Intervention	Truth Intervention
Democrat	253	234	217
Republican	143	134	141

	Female	Male
Democrat	423	281
Republican	219	199



Supplementary Figure 4: Distributions of Obstructionism measure by condition, collapsed across all scenarios, in Experiment 6. Red dots and corresponding numbers are sample means, the boxplot center lines are sample medians. Conditions and scenarios are between subjects.

Experiment 6 “True Values”

In Experiment 6 participants in the intervention conditions were told the true values (i.e. the actual-perceptions) of their out-group and in-group, for the scenario the participant read. Below are those true-values. These values are the mean values, by party and scenario, from the general population sample in Experiment 4.

Supplementary Table 1: True Actual-Perceptions

	Scenario #1	Scenario #2	Scenario #3	Scenario #4	Scenario #5
Dem Actual Disliking	73	45	64	31	49
Dem Actual Unacceptable	73	44	63	33	52
Dem Actual Opposition	76	45	65	31	49
Rep Actual Disliking	74	49	67	46	53
Rep Actual Unacceptable	73	50	69	52	55
Rep Actual Opposition	77	54	71	48	57

Supplementary Methods*Supplemental Experiment A: Convenience Sample Direct Replication of Experiment 3*

Supplemental Experiment A is near-identical to Experiment 4 (in the manuscript). It serves as a direct replication of the effects observed in Experiment 4 but whereas Experiment 4 was run using a nationally representative Qualtrics Panel, Supplemental Experiment A was run on Mechanical Turk. It was performed before Experiment 4 was performed, and the data collected in Supplemental Experiment A was used to conduct the power analysis for the preregistration of Experiment 4.

Outside of the sample, the only way that Supplemental Experiment A differed from Experiment 4 was in the placing of the demographic questions. In Supplement Experiment A the demographic questions appeared at the very end of the survey, and asked for participant's gender and age. In Experiment 4, because the experiment utilized demographic quotas, all the demographic questions appeared at the beginning of the survey, and the questions were expanded to include age, gender, ethnicity, education, and income. As such, the addition of the income, ethnicity, and education questions, along with all the demographic questions being at the

beginning rather than end of the survey, constitute the only differences between Supplemental Experiment A and Experiment 4. See the manuscript for details on Experiment 4's design. Below are the summary statistics and results.

Total N = 397
 $M_{\text{age}} = 35.4$, $SD_{\text{age}} = 11.0$
 199 Women, 198 Men
 260 Democrats, 137 Republicans

	Actual-P Condition	Ingroup-P Condition	Meta-P Condition		Female	Male
Democrat	80	96	84	Democrat	132	128
Republican	56	40	41	Republican	67	70

Supplemental Experiment B: Follow Up on Experiment 6

Supplemental Experiment B was an exploratory follow up study with participants who completed Experiment 6. The follow up occurred approximately a week after participants finished Experiment 6. The goal of Supplemental Experiment B was to examine whether the effects observed in Experiment 6, namely the significant reduction of perceived out-group obstructionism in the intervention conditions and moderation of this effect by accuracy, would last for a weeklong period. In short, we found no evidence of the effect of Experiment 6 a week later.

All 1122 participants from Experiment 6 were directly invited (via email through Mechanical Turks interface) to participate in Supplemental Experiment B. We decided a priori that we would attempt to recruit participants for Supplemental Experiment B for a five-day period, at which point we could cease data collection and analyze the data.

Participants, after providing informed consent, provided their political party affiliation, then responded to a general question regarding out-group obstructionism (“Overall, [out-group members] are purposefully obstructing the legislative process”, 1-100 sliding scale, “Strongly Disagree” to “Strongly Agree”). Participants then provided their age and gender, and the study ended. Participants were paid \$0.50.

In total we collected 886 responses. We then matched participants by gender, Mturk ID, and party affiliation at T1 and T2. This resulted in 64 participants being dropped due to a mismatch in reported political party or gender (8 for gender mismatch, 51 for political party mismatch, 5 for both gender and party mismatch). As such our final sample was $N = 822$. Supplemental Experiment B was not preregistered. Below are the summary statistics and results.

Total N = 822
 479 Women, 343 Men
 529 Democrats, 293 Republicans

Participants by Condition at Time 1

	Control	Hypocrisy Intervention	Truth Intervention
Democrat	189	180	160
Republican	102	95	96

	Female	Male
Democrat	325	204
Republican	154	139

Supplementary Analysis

Supplemental Experiment A: Analysis

Mixed-effect beta regression analysis revealed significant differences between all three conditions on all three outcome measures. Actual perceptions were lower than in-group perceptions for opposition ($b = -0.42$, 95% CI = [-0.58,-0.26], OR = 0.66, $z = -5.12$, $P < 0.001$), unacceptability ($b = -0.31$, 95% CI = [-0.48,-0.15], OR = 0.73, $z = -3.68$, $P < 0.001$), and disliking ($b = -0.43$, 95% CI = [-0.60,-0.27], OR = 0.65, $z = -5.28$, $P < 0.001$). In-group perceptions were lower than GMPs for opposition ($b = 0.73$, 95% CI = [0.55,0.90], OR = 2.07, $z = 8.22$, $P < 0.001$), unacceptability ($b = 0.67$, 95% CI = [0.49,0.85], OR = 1.96, $z = 7.34$, $P <$

0.001), and disliking ($b = 0.72$, 95% CI = [0.54,0.89], OR = 2.05, $z = 8.07$, $P < 0.001$). The pairwise post-hoc contrasts between actual-perceptions and GMPs were also significant for opposition ($b = -1.15$, 95% CI = [-1.35,-0.94], OR = 0.32, $t(1969) = -13.17$, $P < 0.001$), unacceptability ($b = -0.98$, 95% CI = [-1.19,-0.77], OR = 0.37, $t(1970) = -10.98$, $P < 0.001$), and disliking ($b = -1.15$, 95% CI = [-1.36,-0.95], OR = 0.32, $t(1972) = -13.15$, $P < 0.001$). These results directly replicate the findings from Experiment 4. All these models are main-effects only models, as party-accuracy never significantly interacted with condition for any of the DVs (also replicating the findings from Experiment 4).

Supplemental Experiment B: Analysis

To investigate the effect of the T1 intervention and accuracy with T2 perceived obstructionism ($M = 75.69$, $SD = 21.14$), we utilized a multiple regression framework, with T2 obstructionism as the dependent variable regressed onto T1 condition (control, truth intervention, hypocrisy intervention), T1 accuracy (continuous), and an interaction of T1 condition and accuracy.

We find no evidence that obstructionism differed from control in either the truth intervention ($b = 2.43$, 95% CI = [-2.54,7.42], $t(816) = 0.96$, $P = 0.34$) or hypocrisy intervention ($b = -1.74$, 95% CI = [-6.77,3.30], $t(816) = -0.68$, $P = 0.50$), nor was there a significant interaction of T1 accuracy with the truth intervention ($b = -0.07$, 95% CI = [-0.22,0.09], $t(816) = -0.83$, $P = 0.41$) or hypocrisy intervention ($b = 0.08$, 95% CI = [-0.08,0.25], $t(816) = 1.03$, $P = 0.30$). There was, however, a positive linear association between T1 accuracy and T2 obstructionism ($r = 0.22$, 95% CI = [0.15,0.28], $t(820) = 6.43$, $P < 0.001$), suggesting that those who were more inaccurate and overly negative in their group meta-perceptions at T1 perceived their out-group as being higher in obstructionism at T2.

Appendix

Phase One Survey Language

(managerial sample)

Instructions for Written Account

“Managers are often confronted with business and organizational decisions related to social responsibility, diversity, and institutional values. Such decisions are often complex, and businesses are increasingly paying attention to questions of social responsibility.

Please describe, in 3-5 sentences, a decision you’ve previously made, on the part of an organization, that related to issues surrounding social responsibility, diversity, and/or institutional values. The decision can have taken place at a current or former organization.

After providing your written account you’ll be asked a few questions about how you believe others would perceive your behavior.

Please do not include any identifying information in your written account. This includes information that could identify you, your coworkers, and your organization.”

[OPEN TEXT BOX]

Meta-Motive and Actual Motive Items

[PIPE IN STORY]

“Imagine someone else—the average person—read your written account and had to rate how much they think the decision was motivated by the following factors (below). Please indicate how you think THEY would rate your motivations.”

[All 7-point Likert, “Not at All” to “Completely”]

- Selflessness
- Compassion
- Loyalty to Company
- Personal Self-Interest
- Desire to Help Others
- Fairness
- Sense of Moral Obligation
- Company Interests
- Conflict Avoidance
- Adherence to Regulations
- Financial Concerns
- Adherence to Personal Values
- Adherence to Company Values

“How ethical would the average person perceive this decision?” (-)

- [“Very Unethical” to “Very Ethical”, 1-7 Likert]

“How socially responsible would the average person perceive this decision?”

- [“Very Responsible” to “Very Irresponsible”, 1-7 Likert]

“How right or wrong would the average person perceive this decision?”

- [“Very Right” to “Very Wrong”, 1-7 Likert]

“Based on this decision, how much would the average person feel [compassion for/sympathy for/moved by] your company?”

- [“Not at All” to “Very Much So”, 1-7 Likert]

“Based on this decision, how would the average person rate your company on the following characteristics?” [All 7-point Likert, “Not at All” to “Completely”]

- Trustworthiness
- Caring
- Socially Responsible
- Good to Work For
- Principled

“When you made the decision you described, how much were you motivated by the following factors?”

[All 7-point Likert, “Not at All” to “Completely”]

- Selflessness
- Compassion
- Loyalty to Company
- Personal Self-Interest
- Desire to Help Others
- Fairness
- Sense of Moral Obligation
- Company Interests
- Conflict Avoidance
- Adherence to Regulations
- Financial Concerns
- Adherence to Personal Values
- Adherence to Company Values

“How ethical do you consider this decision to be?” (-)

- [“Very Unethical” to “Very Ethical”, 1-7 Likert]

“How socially responsible do you consider this decision to be?”

- [“Very Responsible” to “Very Irresponsible”, 1-7 Likert]

“How right or wrong do you consider this decision to be?”

- [“Very Right” to “Very Wrong”, 1-7 Likert]

Explicit Consent

This study is interested in understanding managers’ motives behind important decisions, and how people will judge managers’ decisions. As such, we would like to use your written account in future studies as stimuli for participants to read. Your written account would only be used for scientific purposes and would be completely anonymized.

Do you consent to allow us to use your written account in future research?

- Yes, I consent to allowing my written account to be used in future research
- No, I do not consent to allowing my written account to be used in future research

Personality Items

Interpersonal Reactivity Index: Perspective-Taking and Empathic Concerns Subscale (Davis, 1983)

“Instructions: The following statements inquire about your thoughts and feelings in a variety of situations. For each item, indicate how well it describes you. Read each item carefully before responding. Answer as honestly as you can.”

[Perspective-Taking IRI Subscale (1-7 Likert, “Does not describe me well” to “Describes me very well”)]

- I sometimes find it difficult to see things from the "other guy's" point of view. (-)
- I try to look at everybody's side of a disagreement before I make a decision.
- I sometimes try to understand my friends better by imagining how things look from their perspective.
- If I'm sure I'm right about something, I don't waste much time listening to other people's arguments. (-)
- I believe that there are two sides to every question and try to look at them both.
- When I'm upset at someone, I usually try to "put myself in his shoes" for a while.
- Before criticizing somebody, I try to imagine how I would feel if I were in their place.

[Empathic-Concern IRI Subscale (1-7 Likert, “Does not describe me well” to “Describes me very well”)]

- I often have tender, concerned feelings for people less fortunate than me.
- Sometimes I don't feel very sorry for other people when they are having problems. (-)
- When I see someone being taken advantage of, I feel kind of protective towards them.
- Other people's misfortunes do not usually disturb me a great deal. (-)

- When I see someone being treated unfairly, I sometimes don't feel very much pity for them. (-)
- I am often quite touched by things that I see happen.
- I would describe myself as a pretty soft-hearted person.

Trait-Machiavellianism (Dahling et al., 2009)

“Instructions: Please indicate the extent to which you agree or disagree with each statement below. There are no right or wrong answers, so please give your honest reaction.”

[1-7 Likert, “Strongly Disagree” to “Strongly Agree”]

- I am willing to be unethical if I believe it will help me succeed.
- I am willing to sabotage the efforts of other people if they threaten my own goals.
- I would cheat if there was a low chance of getting caught.
- I believe that lying is necessary to maintain a competitive advantage over others.
- The only good reason to talk to others is to get information that I can use to my benefit.
- I like to give the orders in interpersonal situations.
- I enjoy being able to control the situation.
- I enjoy having control over other people.
- Status is a good sign of success in life.
- Accumulating wealth is an important goal for me.
- I want to be rich and powerful someday.
- People are only motivated by personal gain.
- I dislike committing to groups because I don't trust others.
- Team members backstab each other all the time to get ahead.
- If I show any weakness at work, other people will take advantage of it.
- Other people are always planning ways to take advantage of the situation at my expense

Propensity for Workplace Deviant Behavior (Chen & Tang, 2006)

“When working at a company, indicate how likely you would be to engage in each of the listed behaviors”

(1 “Very Unlikely” to 7 “Very Likely”).

1. Use office supplies, Xerox machine, and stamps for personal purposes.
2. Make personal long-distance phone calls at work.
3. Waste company time surfing on the internet, playing computer games, and socializing.
4. Take no action for shoplifting by customers
5. Take no action for employees who steal cash/merchandise
6. Take no action for the fraudulent charges made by one's company
7. Borrow \$20 from a cash register overnight without asking.
8. Take merchandise and/or cash home.
9. Give merchandise away for free to personal friends.
10. Abuse the company expense accounts and falsify accounting records.
11. Receive gifts, money, and loans (bribery) from others due to one's position and power.
12. Lay off 500 employees to save the company money and increase one's personal bonus.

13. Overcharge customers to increase sales and earn a higher bonus.
14. Give customers “discounts” first and then secretly charge them more money later (bait and switch).
15. Make more money by deliberately not letting clients know about their benefits

Personal Information

Thank you for participating in this survey. Below we ask you for information about yourself and the organization you work for. **This information will never be shared with anyone outside the research team.** If you gave us consent to share your story, your story will never be connected with information that could identify you or your organization.

All questions below are optional. While we would greatly appreciate knowing more about you and the organization you work for, we understand completely if you would prefer to respond to only some or none of the questions below. Whether or not you respond to the questions below will not affect compensation (i.e. the charity lottery).

What is your name?

[Text Box]

What organization were you working for in the story you provided?

[Text box]

How would you describe the industry of the organization above?

[Text box]

Are you still working for the organization above?

- Yes

- No (who are you currently working for?) [Text box]

When you were in the position during the story you provided, what was your role/rank in the organization?

[Text box]

If you would like the research team to contact you regarding the findings of the study, please provide an email address here:

[Text box]

Demographic Questions

What is your age?

[Text Box]

What is your gender?

- Male
- Female
- Non-binary/other

Approximately how many years of industry experience do you have?

[Text Box]

Preferred Charity

“

In appreciation of your time, this study offer compensation in the form of donations to charities chosen by participants. Specifically, all participants will list their preferred charity, and a subset of charities will be randomly selected to each received \$300 USD.

Please enter the name of your chosen charity below. Note, the charity must be recognized as a charitable non-profit organization by the US federal government in order to be eligible to receive the donation. ”

[Open text box]

Free Response

[Piped in story]

“Above is the story you wrote. Thank you for sharing this with us. If there is any else you wish to say or share about your story, or if you have any comments about your experience with the survey, please share them here.

[Open text box]

[SURVEY ENDS]

Phase Two Survey Language

(general population sample)

Managers' Written Account (Participants will read 5 accounts)

Below is a written account from a high-level manager in a for-profit business. In their written account they describe a decision they've previously made on behalf of their organization. *All these accounts are true*, and written by managers who volunteered to participant in this study.

After reading their written account, you will be asked to judge their decision and their motives. The account below is from a manager at a company in the [industry] sector with approximately [number] employees.

[TEXT of manger's written account]

Judgment Items

“How much do you believe this manager was motivated by the following factors?”

[All 7-point Likert, “Not at All” to “Completely”]

- Selflessness
- Compassion
- Loyalty to Company
- Personal Self-Interest
- Desire to Help Others
- Fairness
- Sense of Moral Obligation
- Company Interests
- Conflict Avoidance
- Adherence to Regulations
- Financial Concerns
- Adherence to Personal Values
- Adherence to Company Values

“How ethical was this decision?” (-)

- [“Very Unethical” to “Very Ethical”, 1-7 Likert]

“How socially responsible was this decision?”

- [“Very Responsible” to “Very Irresponsible”, 1-7 Likert]

“Was this decision right or wrong?”

- [“Very Right” to “Very Wrong”, 1-7 Likert]

“How much do you feel [compassion for/sympathy for/moved by] for the company?”

- [“Not at All” to “Very Much”, 1-7 Likert]

“How trustworthy is this company?”

- [“Very Untrustworthy” to “Very Trustworthy”, 1-7 Likert]

“Based on this decision, how much do you feel [compassion for/sympathy for/moved by] your company?”

- [“Not at All” to “Very Much So”, 1-7 Likert]

“Based on this decision, how would do you rate your company on the following characteristics?”
[All 7-point Likert, “Not at All” to “Completely”]

- Trustworthiness
- Caring
- Socially Responsible
- Good to Work For
- Principled

Personality Items

Interpersonal Reactivity Index: Perspective-Taking and Empathic Concerns Subscale (Davis, 1983)

“Instructions: The following statements inquire about your thoughts and feelings in a variety of situations. For each item, indicate how well it describes you. Read each item carefully before responding. Answer as honestly as you can.”

[Perspective-Taking IRI Subscale (1-7 Likert, “Does not describe me well” to “Describes me very well”)]

- I sometimes find it difficult to see things from the "other guy's" point of view. (-)
- I try to look at everybody's side of a disagreement before I make a decision.
- I sometimes try to understand my friends better by imagining how things look from their perspective.
- If I'm sure I'm right about something, I don't waste much time listening to other people's arguments. (-)
- I believe that there are two sides to every question and try to look at them both.
- When I'm upset at someone, I usually try to "put myself in his shoes" for a while.
- Before criticizing somebody, I try to imagine how I would feel if I were in their place.

[Empathic-Concern IRI Subscale (1-7 Likert, “Does not describe me well” to “Describes me very well”)]

- I often have tender, concerned feelings for people less fortunate than me.
- Sometimes I don't feel very sorry for other people when they are having problems. (-)
- When I see someone being taken advantage of, I feel kind of protective towards them.
- Other people's misfortunes do not usually disturb me a great deal. (-)
- When I see someone being treated unfairly, I sometimes don't feel very much pity for them. (-)
- I am often quite touched by things that I see happen.
- I would describe myself as a pretty soft-hearted person.

Trait-Machiavellianism (Dahling et al., 2009)

“Instructions: Please indicate the extent to which you agree or disagree with each statement below. There are no right or wrong answers, so please give your honest reaction.”

[1-7 Likert, “Strongly Disagree” to “Strongly Agree”]

- I am willing to be unethical if I believe it will help me succeed.
- I am willing to sabotage the efforts of other people if they threaten my own goals.
- I would cheat if there was a low chance of getting caught.
- I believe that lying is necessary to maintain a competitive advantage over others.
- The only good reason to talk to others is to get information that I can use to my benefit.
- I like to give the orders in interpersonal situations.
- I enjoy being able to control the situation.
- I enjoy having control over other people.
- Status is a good sign of success in life.
- Accumulating wealth is an important goal for me.
- I want to be rich and powerful someday.
- People are only motivated by personal gain.
- I dislike committing to groups because I don’t trust others.
- Team members backstab each other all the time to get ahead.
- If I show any weakness at work, other people will take advantage of it.
- Other people are always planning ways to take advantage of the situation at my expense

Propensity for Workplace Deviant Behavior (Chen & Tang, 2006)

“When working at a company, indicate how likely you would be to engage in each of the listed behaviors”

(1 “Very Unlikely” to 7 “Very Likely”).

1. Use office supplies, Xerox machine, and stamps for personal purposes.
2. Make personal long-distance phone calls at work.
3. Waste company time surfing on the internet, playing computer games, and socializing.
4. Take no action for shoplifting by customers
5. Take no action for employees who steal cash/merchandise
6. Take no action for the fraudulent charges made by one’s company
7. Borrow \$20 from a cash register overnight without asking.
8. Take merchandise and/or cash home.
9. Give merchandise away for free to personal friends.
10. Abuse the company expense accounts and falsify accounting records.
11. Receive gifts, money, and loans (bribery) from others due to one’s position and power.
12. Lay off 500 employees to save the company money and increase one’s personal bonus.
13. Overcharge customers to increase sales and earn a higher bonus.
14. Give customers “discounts” first and then secretly charge them more money later (bait and switch).
15. Make more money by deliberately not letting clients know about their benefits

Demographic Questions

What is your age?

[Text Box]

What is your gender?

- Male
- Female
- Non-binary/other

Free Response

“Thank you for participating in our study. If you have any comments about your experience with the survey, please share them here.

[Open text box]

[SURVEY ENDS]

References

- Aaker, J., Vohs, K. D., & Mogilner, C. (2010). Nonprofits are seen as warm and for-profits as competent: Firm stereotypes matter. *Journal of Consumer Research*, 37(August), 224–237. <https://doi.org/10.1086/651566>
- Aggarwal, P., & McGill, A. L. (2012). When brands seem human, do humans act like brands? Automatic behavioral priming effects of brand anthropomorphism. *Journal of Consumer Research*, 39(2), 307–323. <https://doi.org/10.1086/662614>
- Ahn, H.-K., Kim, H. J., & Aggarwal, P. (2014). Helping fellow beings: Anthropomorphized social causes and the role of anticipatory guilt. *Psychological Science*, 25(1), 224–229. <https://doi.org/10.1177/0956797613496823>
- Ames, D., & Fiske, S. (2015). Perceived intent motivates people to magnify observed harms. *Proceedings of the National Academy of Sciences*, 112(12), 3599–3605. <https://doi.org/10.1073/pnas.1501592112>
- Ames, D. R. (2004). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology*, 87(5), 573–585. <https://doi.org/10.1037/0022-3514.87.5.573>
- Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science*, 23(8), 861–868. <https://doi.org/10.1177/0956797611434965>
- Aquino, K., & Reed, A., II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6), 1423–1440. <https://doi.org/10.1037//0022-3514.83.6.1423>
- Ashforth, B. E., & Anand, V. (2003). The normalization of corruption in organizations. *Research in Organizational Behavior*, 25(03), 1–52. [https://doi.org/10.1016/S0191-3085\(03\)25001-2](https://doi.org/10.1016/S0191-3085(03)25001-2)
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Barranti, M., Carlson, E. N., & Côté, S. (2017). How to test questions about similarity in personality and social psychology research: Description and empirical demonstration of response surface analysis. *Social Psychological and Personality Science*, 8(4), 465–475. <https://doi.org/10.1177/1948550617698204>
- Bartels, D. M., Bauman, C. W., Cushman, F., Pizarro, D. A., & McGraw, A. P. (2014). Moral judgments and decision making. In G. Karen & G. Wu (Eds.), *Blackwell Reader of Judgment and Decision Making*.

- Bauman, C. W., Wisneski, D. C., & Skitka, L. J. (2012). Cubist consequentialism: The pros and cons of an agent–patient template for morality. *Psychological Inquiry*, 23(2), 129–133. <https://doi.org/10.1080/1047840X.2012.668014>
- Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, 115(2), 243–267. <https://doi.org/10.1037/0033-2909.115.2.243>
- Bazerman, M. H., & Gino, F. (2012). Behavioral ethics: Toward a deeper understanding of moral judgment and dishonesty. *Annual Review of Law and Social Science*, 8, 85–104. <https://doi.org/10.1146/annurev-lawsocsci-102811-173815>
- Benoit, R. G., & Schacter, D. L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia*, 75, 450–457. <https://doi.org/10.1016/j.neuropsychologia.2015.06.034>
- Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research*, 45(5), 853–885. <https://doi.org/10.1080/00273171.2010.519262>
- Biesanz, J. C., & Human, L. J. (2010). The cost of forming more accurate impressions: Accuracy-motivated perceivers see the personality of others more distinctively but less normatively than perceivers without an explicit goal. *Psychological Science*, 21(4), 589–594. <https://doi.org/10.1177/0956797610364121>
- Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, 41(4), 540–558. <https://doi.org/10.1177/0146167215572134>
- Bolino, M. C. (1999). Citizenship and impression management: Good soldiers or good actors? *Academy of Management Review*, 24(1), 82–98. <https://doi.org/10.2307/259038>
- Bolino, M. C., Varela, J. A., Bande, B., & Turnley, W. H. (2006). The impact of impression-management tactics on supervisor ratings of organizational citizenship behavior. *Journal of Organizational Behavior*, 27(3), 281–297. <https://doi.org/10.1002/job.379>
- Brooks, M., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C., Nielsen, A., Skaug, H., Maechler, M., & Bolker, B. (2017). GlmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2), 378–400.
- Bush, G. W. (2001, September 20). President Bush’s address to a joint session of Congress and the nation. *The Washington Post*. http://www.washingtonpost.com/wp-srv/nation/specials/attacked/transcripts/bushaddress_092001.html

- Capraro, V., & Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making*, 13.
- Carlson, E. N. (2016). Meta-accuracy and relationship quality: Weighing the costs and benefits of knowing what people really think about you. *Journal of Personality and Social Psychology*, 111(2), 250–264. <https://doi.org/10.1037/pspp0000107>
- Carlson, E. N., Furr, R. M., & Vazire, S. (2010). Do we know the first impressions we make? Evidence for idiographic meta-accuracy and calibration of first impressions. *Social Psychological and Personality Science*, 1(1), 94–98. <https://doi.org/10.1177/1948550609356028>
- Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-insight: Do people really know how others see them? *Journal of Personality and Social Psychology*, 101(4), 831–846. <https://doi.org/10.1037/a0024297>
- Carr, A. (2015, June 15). The inside story of Starbucks's race together campaign, no foam. *Fast Company*. <https://www.fastcompany.com/3046890/the-inside-story-of-starbucks-race-together-campaign-no-foam>
- Carson, T. L. (2003). Self-interest and business ethics: Some lessons of the recent corporate scandals. *Journal of Business Ethics*, 43(4), 389–394.
- Caruso, E. M. (2010). When the future feels worse than the past: A temporal inconsistency in moral judgment. *Journal of Experimental Psychology: General*, 139(4), 610.
- Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2016). When minds matter for moral judgment: Intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, 11(3), 476–484. <https://doi.org/10.1093/scan/nsv131>
- Chambers, J. R., Baron, R. S., & Inman, M. L. (2006). Misperceptions in intergroup conflict. *Psychological Science*, 17(1), 38–45. <https://doi.org/10.1111/j.1467-9280.2005.01662.x>
- Chambers, J. R., Epley, N., Savitsky, K., & Windschitl, P. D. (2008). Knowing too much: Using private knowledge to predict how one is viewed by others. *Psychological Science*, 19(6), 542–548.
- Chambers, J. R., & Melnyk, D. (2006). Why do I hate thee? Conflict misperceptions and intergroup mistrust. *Personality and Social Psychology Bulletin*, 32(10), 1295–1311. <https://doi.org/10.1177/0146167206289979>
- Chen, Y.-J., & Tang, T. L.-P. (2006). Attitude toward and propensity to engage in unethical behavior: Measurement invariance across major among university students. *Journal of Business Ethics*, 69(1), 77–93. <https://doi.org/10.1007/s10551-006-9069-6>

- Coffee, J. C. (2005). A theory of corporate scandals: Why the USA and Europe differ. *Oxford Review of Economic Policy*, 21(2), 198–211. <https://doi.org/10.1093/oxrep/gri012>
- Cohen, T. R., & Morse, L. (2014). Moral character: What it is and what it does. *Research in Organizational Behavior*, 34, 43–61. <https://doi.org/10.1016/j.riob.2014.08.003>
- Cooley, E., Payne, B. K., Cipolli, W., Cameron, C. D., Berger, A., & Gray, K. (2017). The paradox of group mind: “People in a group” have more mind than “a group of people”. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0000293>
- Crilly, D., Schneider, S. C., & Zollo, M. (2008). Psychological antecedents to socially responsible behavior. *European Management Review*, 5(3), 175–190. <https://doi.org/10.1057/emr.2008.15>
- Cronbach, L. J. (1955). Processes affecting scores on “understanding of others” and “assumed similarity.” *Psychological Bulletin*, 52(3), 177–193. <https://doi.org/10.1037/h0044919>
- 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>
- Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, 17(3), 273–292. <https://doi.org/10.1177/1088868313495594>
- Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, 6, 97–103. <https://doi.org/10.1016/j.copsyc.2015.06.003>
- Cushman, F., Kumar, V., & Railton, P. (2017). Moral learning: Psychological and philosophical perspectives. *Cognition*, 167, 1–10. <https://doi.org/10.1016/j.cognition.2017.06.008>
- Dahling, J. J., Whitaker, B. G., & Levy, P. E. (2009). The development and validation of a new machiavellianism scale. *Journal of Management*, 35(2), 219–257. <https://doi.org/10.1177/0149206308318618>
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113–126. <https://doi.org/10.1037/0022-3514.44.1.113>
- De Roeck, K., & Delobbe, N. (2012). Do environmental CSR initiatives serve organizations’ legitimacy in the oil industry? Exploring employees’ reactions through organizational identification theory. *Journal of Business Ethics*, 110(4), 397–412. <https://doi.org/10.1007/s10551-012-1489-x>

- De Roeck, K., & Maon, F. (2016). Building the theoretical puzzle of employees' reactions to corporate social responsibility: An integrative conceptual framework and research agenda. *Journal of Business Ethics*. <https://doi.org/10.1007/s10551-016-3081-2>
- Dillon, K. D., & Cushman, F. (2012). Agent, patient ... ACTION! What the dyadic model misses. *Psychological Inquiry*, *23*(2), 150–154. <https://doi.org/10.1080/1047840X.2012.668002>
- Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal*, *36*(6), 37.
- Eisenkraft, N., Elfenbein, H. A., & Kopelman, S. (2017). We know who likes us, but not who competes against us: Dyadic meta-accuracy among work colleagues. *Psychological Science*, *28*(2), 233–241.
- Ember, S. (2015, March 18). Starbucks initiative on race relations draws attacks online. *The New York Times*. <https://www.nytimes.com/2015/03/19/business/starbucks-race-together-shareholders-meeting.html>
- Emler, N. (1990). A social psychology of reputation. *European Review of Social Psychology*, *1*(1), 171–193. <https://doi.org/10.1080/14792779108401861>
- Enders, A. M., & Armaly, M. T. (2018). The differential effects of actual and perceived polarization. *Political Behavior*, *41*(3), 815–839.
- Epley, N., & Tannenbaum, D. (2017). Treating ethics as a design problem. *Behavioral Science & Policy*, *3*(2), 72–84. <https://doi.org/10.1353/bsp.2017.0014>
- Epley, N., & Waytz, A. (2010). Mind perception. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology (5th Ed.)* (pp. 498–541). Wiley. <https://doi.org/10.1002/9780470561119.socpsy001014>
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>
- 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>
- Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. *Journal of Personality and Social Psychology*, *114*(4), 547–571. <https://doi.org/10.1037/pspa0000115>

- Fabrizi, M., Mallin, C., & Michelon, G. (2014). The role of CEO's personal incentives in driving corporate social responsibility. *Journal of Business Ethics*, *124*(2), 311–326. <https://doi.org/10.1007/s10551-013-1864-2>
- Farooq, O., Rupp, D. E., & Farooq, M. (2017). The multiple pathways through which internal and external corporate social responsibility influence organizational identification and multifoci outcomes: The moderating role of cultural and social orientations. *Academy of Management Journal*, *60*(3), 954–985. <https://doi.org/10.5465/amj.2014.0849>
- Finchilescu, G. (2010). Intergroup anxiety in interracial interaction: The role of prejudice and metastereotypes. *Journal of Social Issues*, *66*(2), 334–351. <https://doi.org/10.1111/j.1540-4560.2010.01648.x>
- Fiske, A. P., & Rai, T. S. (2015). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.
- Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, *71*(1), 17–38. <https://doi.org/10.1016/j.jml.2013.10.002>
- Frey, F. E., & Tropp, L. R. (2006). Being seen as individuals versus as group members: Extending research on metaperception to intergroup contexts. *Personality and Social Psychology Review*, *10*(3), 265–280. https://doi.org/10.1207/s15327957pspr1003_5
- Furr, R. M. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality*, *76*(5), 1267–1316. <https://doi.org/10.1111/j.1467-6494.2008.00521.x>
- Gaesser, B., & Schacter, D. L. (2014). Episodic simulation and episodic memory can increase intentions to help others. *Proceedings of the National Academy of Sciences*, *111*(12), 4415–4420. <https://doi.org/10.1073/pnas.1402461111>
- Gaesser, Brendan, DiBiase, H. D., & Kensinger, E. A. (2017). A role for affect in the link between episodic simulation and prosociality. *Memory*, *25*(8), 1052–1062. <https://doi.org/10.1080/09658211.2016.1254246>
- Galinsky, A. D., & Moskowitz, G. B. (2000). Counterfactuals as behavioral primes: Priming the simulation heuristic and consideration of alternatives. *Journal of Experimental Social Psychology*, *36*(4), 384–409. <https://doi.org/10.1006/jesp.1999.1409>
- Giacalone, R. A., Jurkiewicz, C. L., & Deckop, J. R. (2008). On ethics and social responsibility: The impact of materialism, postmaterialism, and hope. *Human Relations*, *61*(4), 483–514. <https://doi.org/10.1177/0018726708091019>
- Gilbert, D. T., & Wilson, T. D. (2007). Propection: Experiencing the future. *Science*, *317*(5843), 1351–1354. <https://doi.org/10.1126/science.1144161>

- Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*, *75*(2), 332–346. <https://doi.org/10.1037/0022-3514.75.2.332>
- Glavas, A. (2016). Corporate social responsibility and organizational psychology: An integrative review. *Frontiers in Psychology*, *7*. <https://doi.org/10.3389/fpsyg.2016.00144>
- Glavas, A., & Godwin, L. N. (2013). Is the perception of 'goodness' good enough? Exploring the relationship between perceived corporate social responsibility and employee organizational identification. *Journal of Business Ethics*, *114*(1), 15–27. <https://doi.org/10.1007/s10551-012-1323-5>
- Goldberg, L. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42.
- Goldenberg, A., Saguy, T., & Halperin, E. (2014). How group-based emotions are shaped by collective emotions: Evidence for emotional transfer and emotional burden. *Journal of Personality and Social Psychology*, *107*(4), 581–596. <https://doi.org/10.1037/a0037462>
- Goldstein, N. J., Vezich, I. S., & Shapiro, J. R. (2014). Perceived perspective taking: When others walk in our shoes. *Journal of Personality and Social Psychology*, *106*(6), 941–960. <https://doi.org/10.1037/a0036395>
- Gond, J.-P., El Akremi, A., Swaen, V., & Babu, N. (2017). The psychological microfoundations of corporate social responsibility: A person-centric systematic review. *Journal of Organizational Behavior*, *38*(2), 225–246. <https://doi.org/10.1002/job.2170>
- 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. (2014). Morality beyond the lab. *Science*, *345*(6202), 1242–1242. <https://doi.org/10.1126/science.1259500>
- Graham, J. (2015). Explaining away differences in moral judgment: Comment on Gray and Keeney (2015). *Social Psychological and Personality Science*, 1–5. <https://doi.org/10.1177/1948550615592242>
- Grappi, S., Romani, S., & Bagozzi, R. P. (2013). Consumer response to corporate irresponsible behavior: Moral emotions and virtues. *Journal of Business Research*, *66*(10), 1814–1821. <https://doi.org/10.1016/j.jbusres.2013.02.002>
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *619*(February), 10–11. <https://doi.org/10.1126/science.1134475>

- Gray, K., & Keeney, J. E. (2015). Disconfirming moral foundations theory on its own terms: Reply to Graham (2015). *Social Psychological and Personality Science*, 1–4. <https://doi.org/10.1177/1948550615592243>
- Gray, Kurt. (2012). The power of good intentions: Perceived benevolence soothes pain, increases pleasure, and improves taste. *Social Psychological and Personality Science*, 3(5), 639–645. <https://doi.org/10.1177/1948550611433470>
- Gray, Kurt, Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, 143(4), 1600–1615. <https://doi.org/10.1037/a0036149>
- Gray, Kurt, & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96(3), 505–520. <https://doi.org/10.1037/a0013748>
- Gray, Kurt, & 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, Kurt, 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. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, 167, 66–77. <https://doi.org/10.1016/j.cognition.2017.03.004>
- Groening, C., & Kanuri, V. K. (2013). Investor reaction to positive and negative corporate social events. *Journal of Business Research*, 66(10), 1852–1860. <https://doi.org/10.1016/j.jbusres.2013.02.006>
- Haidt, J., & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology (5th Ed.)* (pp. 797–832). Wiley. <https://doi.org/10.1002/9780470561119.socpsy002022>
- Haidt, Jonathan. (2012). *The righteous mind: Why good people are divided by religion and politics*. Pantheon.
- Haran, U. (2013). A person-organization discontinuity in contract perception: Why corporations can get away with breaking contracts but individuals cannot. *Management Science*, 59(12), 2837–2853. <https://doi.org/10.1287/mnsc.2013.1745>
- Helzer, E. G., Furr, R. M., Hawkins, A., Barranti, M., Blackie, L. E. R., & Fleeson, W. (2014). Agreement on the perception of moral character. *Personality & Social Psychology Bulletin*, 40(12), 1698–1710. <https://doi.org/10.1177/0146167214554957>

- Hibbert, P., & Cunliffe, A. (2015). Responsible management: Engaging moral reflexive practice through threshold concepts. *Journal of Business Ethics*, *127*(1), 177–188. <https://doi.org/10.1007/s10551-013-1993-7>
- Hoch, S. J., & Research, C.-D. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality & Social Psychology*, *53*(2), 221–234.
- Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. <https://doi.org/10.1126/science.1251560>
- Humphreys, M., & Brown, A. D. (2008). An analysis of corporate social responsibility at credit line: A narrative approach. *Journal of Business Ethics*, *80*(3), 403–418.
- Insko, C. A., Schopler, J., Hoyle, R. H., Dardis, G. J., & Graetz, K. A. (1990). Individual-group discontinuity as a function of fear and greed. *Journal of Personality and Social Psychology*, *58*(1), 68–79. <https://doi.org/10.1037/0022-3514.58.1.68>
- Jago, A. (2019). Algorithms and authenticity. *Academy of Management Discoveries*, *5*(1).
- Jago, A. S., & Laurin, K. (2017). Corporate personhood: Lay perceptions and ethical consequences. *Journal of Experimental Psychology: Applied*, *23*(1), 100–113. <https://doi.org/10.1037/xap0000106>
- Jago, A. S., & Pfeffer, J. (2018). Organizations appear more unethical than individuals. *Journal of Business Ethics*. <https://doi.org/10.1007/s10551-018-3811-8>
- Kahneman, D. (2011). *Thinking, fast and slow* (1st Ed.). Farrar, Straus and Giroux.
- Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, *102*(3), 13.
- Kenny, D. A., & DePaulo, B. M. (1993). Do people know how others view them? An empirical and theoretical account. *Psychological Bulletin*, *114*(1), 145–161.
- Kervyn, N., Fiske, S. T., & Malone, C. (2012). Brands as intentional agents framework: How perceived intentions and ability can map brand perception. *Journal of Consumer Psychology*, *22*(2), 166–176. <https://doi.org/10.3851/IMP2701.Changes>
- Klein, O., & Azzi, A. E. (2001). The strategic confirmation of meta-stereotypes: How group members attempt to tailor an out-group's representation of themselves. *British Journal of Social Psychology*, *40*(2), 279–293. <https://doi.org/10.1348/014466601164759>
- Klotz, A. C., & Bolino, M. C. (2013). Citizenship and counterproductive work Behavior: A moral licensing view. *Academy of Management Review*, *38*(2), 292–306. <https://doi.org/10.5465/amr.2011.0109>

- Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, 9(8), 357–359. <https://doi.org/10.1016/j.tics.2005.06.011>
- Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7(1), 67–83. <https://doi.org/10.1007/s11097-007-9066-y>
- Koehn, D., & Goranova, M. (2016). Do investors see value in ethically sound CEO apologies? Investigating stock market reaction to CEO apologies. *Journal of Business Ethics*. <https://doi.org/10.1007/s10551-016-3301-9>
- Kteily, N., Hodson, G., & Bruneau, E. (2016). They see us as less than human: Metadehumanization predicts intergroup conflict via reciprocal dehumanization. *Journal of Personality and Social Psychology*, 110(3), 343–370. <https://doi.org/10.1037/pspa0000044>
- Ku, G., Wang, C. S., & Galinsky, A. D. (2015). The promise and perversity of perspective-taking in organizations. *Research in Organizational Behavior*, 35(November), 79–102. <https://doi.org/10.1016/j.riob.2015.07.003>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Laing, R. D., Phillipson, H., & Lee, A. R. (1966). *Interpersonal perception: A theory and method of research*. Springer.
- Lange, D., & Washburn, N. T. (2012). Understanding attributions of corporate social irresponsibility. *Academy of Management Review*, 37(2), 300–326. <https://doi.org/10.5465/amr.2010.0522>
- Lange, Donald, Lee, P. M., & Dai, Y. (2011). Organizational reputation: A review. *Journal of Management*, 37(1), 153–184. <https://doi.org/10.1177/0149206310390963>
- Lau, T., Morewedge, C. K., & Cikara, M. (2016). Overcorrection for social-categorization information moderates impact bias in affective forecasting. *Psychological Science*, 27(10), 1340–1351.
- Lee, K., & Ashton, M. C. (2004). Psychometric Properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, 39(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8
- Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, 4(3), 279–286. <https://doi.org/10.1038/s41562-019-0766-4>

- Lees, J., & Gino, F. (2017). Is the moral domain unique? A social influence perspective for the study of moral cognition. *Social and Personality Psychology Compass*, 11(8), e12327. <https://doi.org/10.1111/spc3.12327>
- Lenth, R. (2019). *emmeans: Estimated marginal means, aka least-squares means*. <https://CRAN.R-project.org/package=emmeans>
- Lickel, B., Hamilton, D. L., & Sherman, S. J. (2001). Elements of a lay theory of groups: Types of groups, relational styles, and the perception of group entitativity. *Personality and Social Psychology Review*, 5(2), 129–140. https://doi.org/10.1207/S15327957PSPR0502_4
- Logan, N. (2016). The Starbucks race together initiative: Analyzing a public relations campaign with critical race theory. *Public Relations Inquiry*, 5(1), 93–113. <https://doi.org/10.1177/2046147X15626969>
- Lüdecke, D. (2019). *sjPlot: Data visualization for statistics in social science*. <https://doi.org/10.5281/zenodo.1308157>
- Merskin, D. (2004). The construction of arabs as enemies: Post-September 11 discourse of George W. Bush. *Mass Communication and Society*, 7(2), 157–175. https://doi.org/10.1207/s15327825mcs0702_2
- Miller, D. T., & Nelson, L. D. (2002). Seeing approach motivation in the avoidance behavior of others: Implications for an understanding of pluralistic ignorance. *Journal of Personality and Social Psychology*, 83(5), 1066–1075. <https://doi.org/10.1037//0022-3514.83.5.1066>
- Monroe, A. E., Guglielmo, S., & Malle, B. F. (2012). Morality goes beyond mind perception. *Psychological Inquiry*, 23(2), 179–184. <https://doi.org/10.1080/1047840X.2012.668271>
- Moore, C. (2008). Moral disengagement in processes of organizational corruption. *Journal of Business Ethics*, 80(1), 129–139. <https://doi.org/10.1007/s10551-007-9447-8>
- Moore, C. (2015). Moral disengagement. *Current Opinion in Psychology*, 6, 199–204. <https://doi.org/10.1016/j.copsyc.2015.07.018>
- Moore, C., & Gino, F. (2013). Ethically adrift: How others pull our moral compass from true North, and how we can fix it. *Research in Organizational Behavior*, 33, 53–77.
- Moore, D. A., Tenney, E., & Haran, U. (2010). Overprecision in judgment. In *Handbook of Judgment and Decision Making*. Wiley.
- Murphy, P. E., & Schlegelmilch, B. B. (2013). Corporate social responsibility and corporate social irresponsibility: Introduction to a special topic section. *Journal of Business Research*, 66(10), 1807–1813. <https://doi.org/10.1016/j.jbusres.2013.02.001>

- Niemi, L., & Young, L. (2014). Blaming the victim in the case of rape. *Psychological Inquiry*, 25(2), 230–233. <https://doi.org/10.1080/1047840X.2014.901127>
- Niemi, L., & Young, L. (2016). When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin*, 42(9), 1227–1242. <https://doi.org/10.1177/0146167216653933>
- Nyhan, B., & Reifler, J. (2018). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, 29(2), 222–244.
- Öberseder, M., Schlegelmilch, B. B., & Murphy, P. E. (2013). CSR practices and consumer perceptions. *Journal of Business Research*, 66(10), 1839–1851. <https://doi.org/10.1016/j.jbusres.2013.02.005>
- O’Leary, C., & Pangemanan, G. (2007). The effect of groupwork on ethical decision-making of accountancy students. *Journal of Business Ethics*, 75(3), 215–228. <https://doi.org/10.1007/s10551-006-9248-5>
- Palazzo, G., & Basu, K. (2007). The ethical backlash of corporate branding. *Journal of Business Ethics*, 73(4), 333–346. <https://doi.org/10.1007/s10551-006-9210-6>
- Panagopoulos, N. G., Rapp, A. A., & Vlachos, P. A. (2016). I think they think we are good citizens: Meta-perceptions as antecedents of employees’ reactions to corporate social responsibility. *Journal of Business Research*, 69(8), 2781–2790. <https://doi.org/10.1016/j.jbusres.2015.11.014>
- Patil, I., & Powell, C. (2018). *ggstatsplot: “Ggplot2” based plots with statistical details*. <https://doi.org/10.5281/zenodo.2074621>
- Pearce, C. L., & Manz, C. C. (2011). Leadership centrality and corporate social ir-responsibility (CSIR): The potential ameliorating effects of self and shared leadership on CSIR. *Journal of Business Ethics*, 102(4), 563–579. <https://doi.org/10.1007/s10551-011-0828-7>
- Pemberton, M. B., Insko, C. A., & Schopler, J. (1996). Memory for and experience of differential competitive behavior of individuals and groups. *Journal of Personality & Social Psychology*, 71(5), 14.
- Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, 39(6), 653–660. [https://doi.org/10.1016/S0022-1031\(03\)00041-6](https://doi.org/10.1016/S0022-1031(03)00041-6)
- Plitt, M., Savjani, R. R., & Eagleman, D. M. (2015). Are corporations people too? The neural correlates of moral judgments about companies and individuals. *Social Neuroscience*, 10(2), 113–125. <https://doi.org/10.1080/17470919.2014.978026>

- Pronin, E. (2008). How we see ourselves and how we see others. *Science*, 320(5880), 1177–1180.
- Puzakova, M., Kwak, H., & Rocereto, J. F. (2013). When humanizing brands goes wrong: The detrimental effect of brand anthropomorphization amid product wrongdoing. *Journal of Marketing*, 77(May), 81–100.
- Rai, T. S., & Diermeier, D. (2015). Corporations are cyborgs: Organizations elicit anger but not sympathy when they can think but cannot feel. *Organizational Behavior and Human Decision Processes*, 126(1), 18–26. <https://doi.org/10.1016/j.obhdp.2014.10.001>
- Railton, P. (2016). Morality and prospection. In M. E. P. Seligman, P. Railton, R. F. Baumeister, & C. Sripada (Eds.), *Homo Prospectus* (pp. 225–280). Oxford University Press.
- Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. *Cognition*, 167, 172–190. <https://doi.org/10.1016/j.cognition.2016.08.015>
- Rayton, B. A., Brammer, S. J., & Millington, A. I. (2015). Corporate social performance and the psychological contract. *Group & Organization Management*, 40(3), 353–377.
- Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional attribution: Multiple inferences about motive-related traits. *Journal of Personality and Social Psychology*, 86(4), 530–544. <https://doi.org/10.1037/0022-3514.86.4.530>
- Revelle, W. (2018). *psych: Procedures for psychological, psychometric, and personality research*. <https://CRAN.R-project.org/package=psych>
- Reynolds, S. J., & Ceranic, T. L. (2007). The effects of moral judgment and moral identity on moral behavior: An empirical examination of the moral individual. *Journal of Applied Psychology*, 92(6), 1610–1624. <https://doi.org/10.1037/0021-9010.92.6.1610>
- Reynolds, S. J., Leavitt, K., & DeCelles, K. A. (2010). Automatic ethics: The effects of implicit assumptions and contextual cues on moral behavior. *Journal of Applied Psychology*, 95(4), 752–760. <https://doi.org/10.1037/a0019411>
- Reynolds, S. J., Owens, B. P., & Rubenstein, A. L. (2012). Moral stress: Considering the nature and effects of managerial moral uncertainty. *Journal of Business Ethics*, 106(4), 491–502. <https://doi.org/10.1007/s10551-011-1013-8>
- Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review*, 9(1), 32–47. https://doi.org/10.1207/s15327957pspr0901_3
- Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: “Naive realism” in intergroup perception and conflict. *Journal of Personality and Social Psychology*, 68(3), 404–417. <https://doi.org/10.1037/0022-3514.68.3.404>

- Rogers, T., & Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, 2(5), 335–342. <https://doi.org/10.1038/s41562-018-0328-1>
- Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, 74, 24–37. <https://doi.org/10.1016/j.jesp.2017.08.003>
- Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology*, 69, 44–58. <https://doi.org/10.1016/j.jesp.2016.09.007>
- Rupp, D. E., & Mallory, D. B. (2015). Corporate social responsibility: Psychological, person-centric, and progressing. *Annual Review of Organizational Psychology and Organizational Behavior*, 2(1), 211–236. <https://doi.org/10.1146/annurev-orgpsych-032414-111505>
- Saguy, T., & Kteily, N. (2011). Inside the opponent's head: Perceived losses in group position predict accuracy in metaperceptions between groups. *Psychological Science*, 22(7), 951–958.
- Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, 8(9), 657–661. <https://doi.org/10.1038/nrn2213>
- Schein, C., & Gray, K. (2014). The prototype model of blame: Freeing moral cognition from linearity and little boxes. *Psychological Inquiry*, 25(2), 236–240. <https://doi.org/10.1080/1047840X.2014.901903>
- Schein, C., & Gray, K. (2016). Moralization and harmification: The dyadic loop explains how the innocuous becomes harmful and wrong. *Psychological Inquiry*, 7965(July), 62–65. <https://doi.org/10.1080/1047840X.2016.1111121>
- Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, 1–39. <https://doi.org/10.1177/1088868317698288>
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5), 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science*, 8(2), 119–141. <https://doi.org/10.1177/1745691612474317>

- Semmann, D., Krambeck, H.-J., & Milinski, M. (2005). Reputation is valuable within and outside one's own social group. *Behavioral Ecology and Sociobiology*, *57*(6), 611–616. <https://doi.org/10.1007/s00265-004-0885-3>
- Sezer, O., Gino, F., & Bazerman, M. H. (2015). Ethical blind spots: Explaining unintentional unethical behavior. *Current Opinion in Psychology*, *6*, 77–81. <https://doi.org/10.1016/j.copsyc.2015.03.030>
- Shanock, L. R., Baran, B. E., Gentry, W. A., Pattison, S. C., & Heggestad, E. D. (2010). Polynomial regression with response surface analysis: A powerful approach for examining moderation and overcoming limitations of difference scores. *Journal of Business and Psychology*, *25*(4), 543–554. <https://doi.org/10.1007/s10869-010-9183-4>
- Sigelman, L., & Tuch, S. A. (1997). Metastereotypes: Blacks' perceptions of whites' stereotypes of blacks. *Public Opinion Quarterly*, *61*(1, Special Issue on Race), 87. <https://doi.org/10.1086/297788>
- Skarmeas, D., & Leonidou, C. N. (2013). When consumers doubt, Watch out! The role of CSR skepticism. *Journal of Business Research*, *66*(10), 1831–1838. <https://doi.org/10.1016/j.jbusres.2013.02.004>
- Sluss, D. M., & Ashforth, B. E. (2008). How relational and organizational identification converge: Processes and conditions. *Organization Science*, *19*(6), 807–823.
- Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. <https://doi.org/10.1037/1082-989X.11.1.54>
- Solomon, B. C., & Vazire, S. (2016). Knowledge of identity and reputation: Do people have knowledge of others' perceptions? *Journal of Personality and Social Psychology*, *111*(3), 341–366. <https://doi.org/10.1037/pspi0000061>
- Stahl, G. K., & Sully de Luque, M. (2014). Antecedents of responsible leader behavior: A research synthesis, conceptual framework, and agenda for future research. *Academy of Management Perspectives*, *28*(3), 235–254. <https://doi.org/10.5465/amp.2013.0126>
- Starbucks Corporation. (2015a). *What 'race together' means for Starbucks partners and customers*. <https://news.starbucks.com/news/what-race-together-means-for-starbucks-partners-and-customers>
- Starbucks Corporation. (2015b). *A letter from Howard Schultz to Starbucks partners regarding race together*. <https://news.starbucks.com/news/a-letter-from-howard-schultz-to-starbucks-partners-regarding-race-together>

- Stern, C., & Kleiman, T. (2015). Know thy outgroup: Promoting accurate judgments of political attitude differences through a conflict mindset. *Social Psychological and Personality Science*, 6(8), 950–958. <https://doi.org/10.1177/1948550615596209>
- Stevens, J. M., Kevin Steensma, H., Harrison, D. A., & Cochran, P. L. (2005). Symbolic or substantive document? The influence of ethics codes on financial executives' decisions. *Strategic Management Journal*, 26(2), 181–195. <https://doi.org/10.1002/smj.440>
- Stroessner, S. J., & Dweck, C. S. (2015). Inferring group traits and group goals. In S. J. Stroessner & J. W. Sherman (Eds.), *Social Perception: From Individuals to Groups* (pp. 177–196). Psychology Press.
- Strohming, N., & Nichols, S. (2014). The essential moral self. *Cognition*, 131(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Sunstein, C. R. (2002). Why they hate us: The role of social dynamics. *Harvard Journal of Law & Public Policy*, 25, 429–440.
- Swanson, D. L. (1995). Addressing a theoretical problem by reorienting the corporate social performance model. *The Academy of Management Review*, 20(1), 43. <https://doi.org/10.2307/258886>
- Sweetin, V. H., Knowles, L. L., Summey, J. H., & McQueen, K. S. (2013). Willingness-to-punish the corporate brand for corporate social irresponsibility. *Journal of Business Research*, 66(10), 1822–1830. <https://doi.org/10.1016/j.jbusres.2013.02.003>
- Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, 47(6), 1249–1254. <https://doi.org/10.1016/j.jesp.2011.05.010>
- Tenbrunsel, A. E., Diekmann, K. A., Wade-Benzoni, K. A., & Bazerman, M. H. (2010). The ethical mirage: A temporal explanation as to why we are not as ethical as we think we are. *Research in Organizational Behavior*, 30(C), 153–173. <https://doi.org/10.1016/j.riob.2010.08.004>
- Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research*, 17(2), 223–236. <https://doi.org/10.1023/B:SORE.0000027411.35832.53>
- Teper, R., Inzlicht, M., & Page-Gould, E. (2011). Are we more moral than we think? Exploring the role of affect in moral behavior and moral forecasting. *Psychological Science*, 22(4), 553–558. <https://doi.org/10.1177/0956797611402513>
- Treviño, L. K., den Nieuwenboer, N. A., & Kish-Gephart, J. J. (2014). (Un)ethical behavior in organizations. *Annual Review of Psychology*, 65(1), 635–660. <https://doi.org/10.1146/annurev-psych-113011-143745>

- Turker, D. (2009). Measuring corporate social responsibility: A scale development study. *Journal of Business Ethics*, 85(4), 411–427. <https://doi.org/10.1007/s10551-008-9780-6>
- Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, 4(6), 13.
- Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, 98(2), 281–300. <https://doi.org/10.1037/a0017908>
- Vazire, S., & Carlson, E. N. (2010). Self-knowledge of personality: Do people know themselves. *Social and Personality Psychology Compass*, 4(8), 605–620.
- Vazire, S., & Carlson, E. N. (2011). Others sometimes know us better than we know ourselves. *Current Directions in Psychological Science*, 20(2), 104–108. <https://doi.org/10.1177/0963721411402478>
- Vlachos, P. A., Panagopoulos, N. G., & Rapp, A. A. (2013). Feeling good by doing good: Employee CSR-induced attributions, job satisfaction, and the role of charismatic leadership. *Journal of Business Ethics*, 118(3), 577–588. <https://doi.org/10.1007/s10551-012-1590-1>
- Vlachos, P. A., Theotokis, A., & Panagopoulos, N. G. (2010). Sales force reactions to corporate social responsibility: Attributions, outcomes, and the mediating role of organizational trust. *Industrial Marketing Management*, 39(7), 1207–1218. <https://doi.org/10.1016/j.indmarman.2010.02.004>
- Vorauer, J. D., Hunter, A., Main, K., & Roy, S. (2000). Meta-stereotype activation: Evidence from indirect measures for specific evaluative concerns experienced by members of dominant groups in intergroup interaction. *Journal of Personality and Social Psychology*, 78(4), 690–707. <https://doi.org/10.1037/0022-3514.78.4.690>
- Vorauer, J. D., Main, K. J., & O’Connell, G. B. (1998). How do individuals expect to be viewed by members of lower status groups? Content and implications of meta-stereotypes. *Journal of Personality & Social Psychology*, 75(4), 21.
- Wang, S., Gao, Y., Hodgkinson, G. P., Rousseau, D. M., & Flood, P. C. (2015). Opening the black box of CSR decision making: A policy-capturing study of charitable donation decisions in china. *Journal of Business Ethics*, 128(3), 665–683. <https://doi.org/10.1007/s10551-014-2123-x>
- Waytz, A., Young, L. L., & Ginges, J. (2014). Motive attribution asymmetry for love vs. Hate drives intractable conflict. *Proceedings of the National Academy of Sciences*, 111(44), 15687–15692. <https://doi.org/10.1073/pnas.1414146111>

- Waytz, Adam, Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. <https://doi.org/10.1016/j.jesp.2014.01.005>
- Waytz, Adam, & Young, L. (2012). The group-member mind trade-off: Attributing mind to groups versus group members. *Psychological Science*, *23*(1), 77–85. <https://doi.org/10.1177/0956797611423546>
- Weaver, G. R., Trevino, L. K., & Cochran, P. L. (1999). Integrated and decoupled corporate social performance: Management commitments, external pressures, and corporate ethics practices. *Academy of Management Journal*, *42*(5), 539–552. <https://doi.org/10.2307/256975>
- West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, *118*(2), 357–378. <https://doi.org/10.1037/a0022936>
- Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science*, *10*(2), 145–158. <https://doi.org/10.1177/1745691615569849>
- Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., & Schopler, J. (2003). Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect. *Psychological Bulletin*, *129*(5), 698–722. <https://doi.org/10.1037/0033-2909.129.5.698>
- Wood, D., & Furr, R. M. (2016). The correlates of similarity estimates are often misleadingly positive: The nature and scope of the problem, and some solutions. *Personality and Social Psychology Review*, *20*(2), 79–99. <https://doi.org/10.1177/1088868315581119>
- Wood, D. J. (1991). Corporate social performance revisited. *The Academy of Management Review*, *16*(4), 691. <https://doi.org/10.2307/258977>
- Yang, J., Ji, H., & O’Leary, C. (2017). Group ethical decision making process in Chinese business: Analysis From social decision scheme and cultural perspectives. *Ethics & Behavior*, *27*(3), 201–220. <https://doi.org/10.1080/10508422.2016.1157690>
- Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240. <https://doi.org/10.1073/pnas.0701408104>
- Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214. <https://doi.org/10.1016/j.cognition.2011.04.005>
- Zakaria, F. (2001, October 14). The politics of rage: Why do they hate us? *Newsweek*. <https://www.newsweek.com/politics-rage-why-do-they-hate-us-154345>

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19*(4), 399–404. <https://doi.org/10.1111/j.1467-9280.2008.02099.x>

Zhang, T., Gino, F., & Bazerman, M. H. (2014). Morality rebooted: Exploring simple fixes to our moral bugs. *Research in Organizational Behavior, 34*, 63–79. <https://doi.org/10.1016/j.riob.2014.10.002>