How can leaders foster trust when making decisions under risk?

Charles A. Dorison

McDonough School of Business

Georgetown University

Abstract

Leaders must make sound judgments and decisions while also maintaining trust from key constituencies. However, constituents regularly distrust leaders who follow value-maximizing decision processes. This tension represents a key leadership challenge. In the present work, I examine whether leaders can predict these reputational costs and test communication strategies they can use to overcome them. I do so in the context of loss-gain framing effects on risk preferences: the robust and widely influential tendency for risk preferences to shift depending on whether outcomes are described as losses or gains. I show that leaders (here, law enforcement executives) correctly anticipate incurring a trust penalty for making decisions consistent across different framings. Moreover, I show that this penalty can be mitigated by expressing learning goals (but not by making consistent decisions or simply explaining framing effects). Across four experiments, I test a conceptual model that explores the central yet complex role of biased risk preferences of both leaders and constituents as a driving mechanism for these results. As a complement to approaches focused on individual-level training, the present research suggests a necessary focus on identifying organizational norms and developing communication strategies to help leaders maintain trust when making effective judgments and decisions. These contributions apply not only for decision making under risk, but also for a host of judgmental biases central to managerial decision making.

Keywords: Leadership, managerial decision making, trust, communication, risk taking

Leaders must make sound judgments and decisions while also maintaining trust from key constituencies (Bazerman & Moore, 2012; Galinsky & Schweitzer, 2015). However, leaders who employ processes demonstrated to improve decision quality often suffer negative reputational consequences. For example, constituents distrust leaders who disregard irrelevant decision frames when making risky decisions (Dorison & Heller, 2022). The same reputational penalties can apply when leaders use Bayesian judgments (Cao, Kleiman-Weiner, & Banaji, 2017), calibrate their confidence (Anderson, Brion, Moore, & Kennedy, 2012; Tenney et al., 2019), avoid costly escalation of commitment (Brockner, 1992; Dorison, Umphres, & Lerner, 2021; Kanodia, 1989), admit moral nuance (Huppert, Herzog, Landy, & Levine, 2023), allocate scarce resources efficiently (Everett, Pizarro, & Crockett, 2016), seek advice from independent advisors (Blunden et al., 2019), consistently follow rules (White, Levine, & Kristal, 2023), and change their minds in the face of evidence (Kreps, Laurin, & Merritt, 2017). Managing the reputational costs systematically associated with high-quality decision processes thus represents a major challenge for leaders.

In the present work, I investigate leaders' abilities to anticipate the reputational costs of following high-quality decision processes and examine strategies they can use to overcome these costs. I do so in the decision context of loss-gain framing effects on risk preferences: the robust and widely influential tendency for risk preferences to shift depending on whether outcomes are described as gains or losses (Kahneman & Tversky, 1979; Ruggeri et al., 2020; Tversky & Kahneman, 1981). Such framing effects are pervasive in diverse domains including buying insurance (Hershey & Schoemaker, 1980), saving for retirement (Benartzi & Thaler, 1995), reaching agreement in negotiations (Bazerman, 1983), and trading commodities (Sun & Mellers, 2016). I show that leaders (correctly) anticipate incurring a trust penalty for making decisions consistent across different framings. Moreover, I show that this penalty can be mitigated by expressing learning goals (but not by making consistent decisions or simply explaining framing effects). Across four experiments, I develop and test a conceptual model that explores the central yet complex role of biased risk preferences of both leaders and observers as a driving mechanism for these results.

Given its central role for personal and professional success, I focus my investigation primarily on trust (Arrow, 1974; Dirks & Ferrin, 2002; Kramer, 1999; Mayer, Davis, & Schoorman, 1995). Sometimes called an "important lubricant of the social system" (Arrow, 1974), trust is a central yet multifaceted concept in organizational life. For example, Mayer, Davis, and Schoorman (1995) theorized three distinct precursors to organizational trust: ability, benevolence, and integrity (for related work, see Dirks & Ferrin, 2002; Huppert, Herzog, Landy, & Levine, 2023; Zlatev, 2019). The present work takes a high-level view of trust, testing it across various specifications (e.g., perceived vs. behavioral trust). I measure a mix of different facets of trust and discuss potential similarities and differences—as well as predictions based on the conceptual model—across experiments.

Broadly, the present research contributes to a growing body of research linking behavioral science, communication, and leadership (Moore & Bazerman, 2022). Prior pioneering research has identified the myriad ways in which accountability structures in organizations shape judgment and choice (Lerner & Tetlock, 1999; Tetlock, 2000). Here, I build on this foundation by exploring not only how constituents evaluate leaders' decisions, but also whether leaders can accurately predict such evaluations and how leaders can communicate their decisions more effectively. As a complement to individual-level training, the present research suggests a necessary focus on identifying organizational norms and developing communication strategies to help leaders overcome canonical decision biases. These contributions apply not only for decision making under risk, but also for a host of judgmental biases central to managerial decision making.

Background

Decision making under risk—choices in situations where relevant probabilities are known—has received sustained and interdisciplinary scholarly attention for decades (Caraco, Martindale, & Whittam, 1980; Gigerenzer et al., 1999; Loewenstein, Webber, Hsee, & Welch, 2001; Mishra, 2014; Tversky & Kahneman, 1979). At the intersection of mathematics and economics, researchers made great strides during the first half of the 20th century in developing models of how decision makers ought to navigate

risk (Edwards, 1954; Edwards, Lindman, & Savage, 1963; Friedman & Savage, 1948, 1952; Savage, 1951; von Neumann & Morgenstern, 1944).

More recently, behavioral decision researchers identified systematic deviations from these models. This research typically strips away the social and organizational context to examine the cognitive underpinnings—and fallibility—of human decision making under risk (e.g., Kahneman & Tversky, 1979, 1981).[1] For example, using the benchmark of expected utility theory as its foil, prospect theory harnessed insights from cognitive psychology to introduce a descriptively accurate model of how individuals (and leaders) navigate such choices. Prospect theory thus gives rise to a suite of widely studied behavioral tendencies that violate basic assumptions of expected utility theory (for reviews, see Arkes, 1991; Gilovich, Griffin, & Kahneman, 2002; Gilovich & Griffin, 2010; Mercer, 2005; Ruggeri et al., 2020).

Among the most influential and widely studied consequences of prospect theory is the loss-gain framing effect on risk preferences: the tendency for individuals to prefer risk-seeking options when choice alternatives are framed as losses, but to prefer risk-averse options when choice alternatives are framed as gains (e.g., Bazerman, 1983; Benartzi & Thaler, 1995; Chen, Lakshminarayanan, & Santos, 2006; Hershey & Schoemaker, 1980; McNeil et al., 1982; Sun & Mellers, 2016; for large-scale replication across 19 countries and 13 languages, see Ruggeri et al., 2020). Persisting even with large financial stakes and with experienced leaders, such loss-gain framing effects on risk preferences are widely considered suboptimal mistakes that violate core tenets of rationality. For example, Tversky and Kahneman (1981, p. 453) noted that "the dependence of preferences on the formulation of decision problems is a significant concern for the theory of rational choice." More recently, Bazerman and Moore (2009, p. 65) filed a concurring and unqualified opinion: "rational decision makers should be immune to the framing of choices."

[1] Although outside the scope of the present article, a lively research program simultaneously examines the *affective* underpinnings of decision making under risk and uncertainty (Lerner & Keltner, 2001; Loewenstein, Webber, Hsee, & Welch, 2001; Slovic, Finucane, Peters, & MacGregor, 2007).

Most prior empirical research on loss-gain framing effects uses laboratory or online experiments in which participants make anonymous choices (for review, see Ruggeri et al., 2020). While taking an intrapersonal perspective successfully elucidates underlying cognitive mechanisms, this simplifying assumption can leave other important interpersonal variables—critical for leaders and organizations—relatively unexplored.

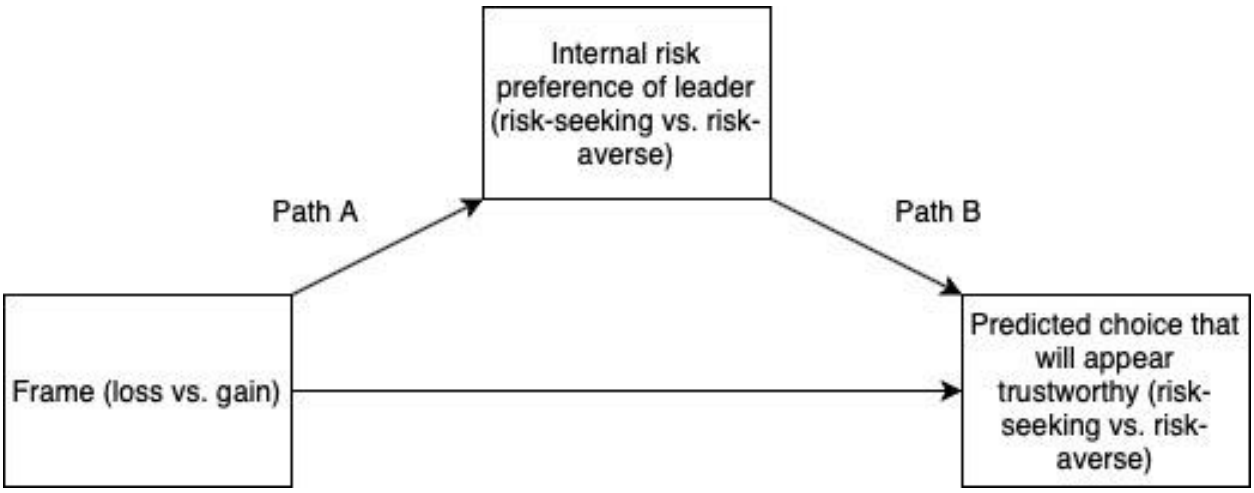
Most relevant for the present research, recent work provides converging evidence that constituents penalize leaders whose risk preferences are *unaffected* by loss-gain framing (Dorison & Heller, 2022). Using a representative sample and financial stakes over thirty times those used in prior work, Supplemental Experiment 1 replicates this finding, confirming that constituents are strongly attuned to decision frames when deciding which leaders to trust. Specifically, using a behavioral measure most relevant to integrity- and benevolence-based trust, participants preferred to place a risk-averse peer in charge of a group resource when the choice options were presented as gains, but preferred to place a risk-seeking peer in charge of a group resource when choice options were presented as losses (for full details, see *SI*).

Given that managing the impressions they leave on others is a primary leadership goal, key questions remain regarding the broader social and organizational context in which loss-gain framing effects on risk preferences operate. For example, can leaders predict these costs? What communication strategies can they use to overcome them? And what psychological mechanisms underpin these predictions and the effectiveness of these strategies? Below, I review prior literature to derive theory-driven hypotheses to answer these questions.

Conceptual development: leader predictions

If observers distrust leaders whose choices are immune to framing effects, are leaders sensitive to these negative evaluations? I drew on research on bias blind spot and naïve realism to predict that they *would* intuit that frames shift how observers evaluate them. Figure 1 presents a two-part conceptual model, which I detail below.

Figure 1. A conceptual model for how leaders predict their reputational incentives. Leaders have biased risk preferences, such that their internal risk preferences are affected by loss-gain frames (Path A). They do not have insight into this bias; instead, they believe they are objective and that reasonable constituents will agree with them (and trust them; Path B).



First, drawing on a robust prior research literature in behavioral decision making (Camerer & Hogarth, 1999; Dawes, Faust, & Meehl, 1989; Enke et al., 2023; Northcraft & Neale, 1987; Massey & Thaler, 2013; McNeil, Pauker, Sox, & Tversky, 1982; Schwitzgebel & Cushman, 2015; Shanteau, 1992; Simmons & Massey, 2012), I predicted that leaders' internal risk preferences would shift as a function of whether choices were framed as losses or gains. That is, I expected even experienced leaders to demonstrate a loss-gain framing effect. Specifically, I expected them to prefer the risk-seeking option when choices were framed as losses and to prefer the risk-averse option when choices were framed as gains. This prediction constitutes "Path A" in the leader prediction model.

More novel to the present investigation is "Path B," which links leaders' (biased) internal risk preferences to their predictions for how constituents will evaluate them. Research on the bias blind spot (Ehrlinger, Gilovich, & Ross, 2005; Pronin, Lin, & Ross, 2002; Scopelliti et al., 2015) and naïve realism

(Ross & Ward, 1995; Schwalbe, Cohen, & Ross, 2016) argue that individuals perceive their own judgments as fair and objective—even when they are not. Consequently, individuals believe that reasonable constituents are likely to agree with their choices. If so, then while leaders' choices shift as a function of the frame in which options are presented, so too will their predictions regarding which choice will be perceived positively by their constituents. Thus, just as leaders' own risk preferences shift, so too will their predictions for which choice will be rewarded. Put another way, while in the loss frame leaders will tend to prefer the risk-seeking option (and believe reasonable others will agree with them, and therefore trust them), in the gain frame leaders will tend to prefer the risk-averse option (and yet still believe reasonable others will agree with them, and therefore trust them). As a result, and despite the egocentric biases inherent in this reasoning style, leaders' reputational forecasts will be sensitive to the frame in which options are presented—and therefore accurate. On a related note, given that attention to observers' perceptions is critical for attaining status and power and moving up a hierarchy (Galinsky & Wang, 2005; Magee & Galinsky, 2008), one might expect that leaders would be *especially* talented at perceiving their reputational incentives, although this relative comparison is not tested in the present manuscript.

Formally, I hypothesized:

Hypothesis 1: Leaders will accurately predict that their reputational incentives shift as a function of decision frames.

Hypothesis 1a: Leaders' internal risk preferences will be sensitive to the frame in which options are presented.

Hypothesis 1b: Leaders' predictions will be driven by their internal (biased) risk preferences.

And yet, three research streams provide suggestive evidence that this model may not hold. A large literature provides converging evidence that individuals systematically misunderstand the consequences of their social interactions and how they are perceived by others (e.g., Epley et al., 2022; Epley & Schroeder, 2014; Gilovich, Medvec, & Savitsky, 2000; Kardas, Schroeder, & O'Brien, 2021; Nisbett & Ross, 1980). A second related line of research on impression (mis)management reveals that individuals

systematically use erroneous and ineffective self-presentation strategies when trying to make positive impressions on others, often without realizing their failures (for reviews reaching this conclusion, see Sezer, 2022; Steinmetz, Sezer, & Sedikides, 2017). For example, employees will often hide their success from peers instead of sharing, even though such hiding yields relational costs (Roberts, Levine, & Sezer, 2020).

Perhaps most relevant, a third related and important stream of research reveals that individuals fail to strategically use decision frames to influence others (Daniels & Zlatev, 2019). In one study, leaders (i.e., participants assigned to present choices to others) failed to successfully use framing effects to shift others' risk preferences. Instead, leaders tended to structure choices using a positive frame, regardless of their goal. Of note, such effects held not only among Stanford undergraduates, but also among professional students (e.g., students at Stanford Law School, Graduate School of Business, and Medical School). Related work in this research program demonstrates the generalizability of these effects across both a variety of influence tools (e.g., default effects) and a variety of other professional samples (e.g., Zlatev, Daniels, Kim, & Neale, 2019; for dissenting opinions, see Jung, Sun, & Nelson, 2019; McKenzie, Leong, & Sher, 2021). To the extent that individuals systematically fail to understand the effects of decision frames on others' choices, it seems reasonable to infer that they might also systematically fail to understand the effect of decision frames on how others evaluate *their* choices.

These three research literatures are most relevant to Path B in the conceptual model (i.e., linking leaders' internal risk preferences to their perceived reputational incentives). I empirically parse these different pathways in Experiment 1 to directly test these possibilities.

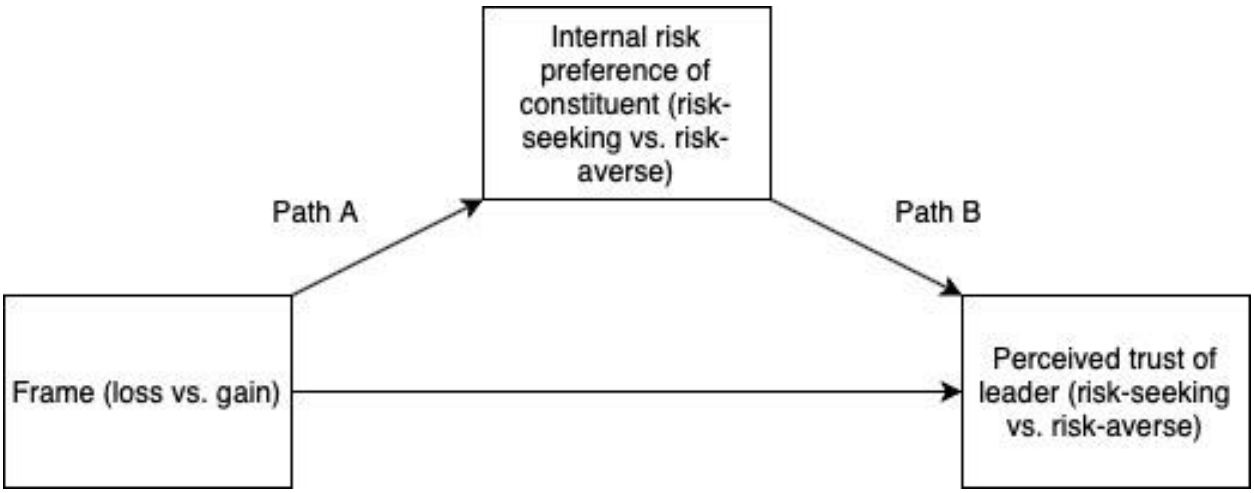
Conceptual development: leader communication strategies

Leaders face a tension between following sound decision processes and fostering trust from key constituencies. An impressive prior literature has focused on increasing the probability of the former, with significant investment in training exercises and strategies to improve decision processes (e.g., Morewedge et al., 2015; Larrick, Morgan, & Nisbett, 1990; Sellier, Scopelliti, & Morewedge, 2019; for review, see Larrick, 2004). But decisions are not made and evaluated in isolation. Instead, leaders can often explain,

justify, and defend their choices to their constituencies. If constituents penalize leaders who have consistent risk preferences (and leaders can predict these penalties), how can leaders defend their risk choices, especially unpopular ones? Relatively little research acknowledges the reputational costs of making unbiased decisions and develops strategies to overcome them.

Figure 2 presents a two-part conceptual model of how leaders can communicate their decisions to shift constituents' evaluations. This model shares deep theoretical foundations with the leader prediction model. First, as with the leader prediction model, it begins with the hypothesis that constituents themselves have biased risk preferences (Path A). Second, again building on research on bias blind spot and naïve realism, it proposes that constituents distrust leaders whose risk preferences diverge from their own (Path B). This model thus sets the stage for two distinct types of strategies leaders could use to maintain trust when making decisions under risk, detailed below.

Figure 2. A conceptual model of how constituents evaluate leaders based on the frame and the leader's choice. Constituents hold biased risk preferences, and in turn reward leaders whose risk preferences align with their own. Leaders can either attempt to shift their constituents' risk preferences (i.e., reduce the amount of disagreement that emerges if they have unbiased risk preferences; Path A) or attempt to reduce the reputational penalties from disagreement (Path B).



One set of strategies could focus on *shifting constituents' risk preferences* (i.e., a strategy focused on Path A). To do so, leaders could maintain consistent risk preferences for multiple decisions over time—thus signaling the correct course of action to their constituents through social learning. Of note, prior research examined constituents' evaluations of a leader who made the risk-seeking *or* risk-averse choice in the loss *or* gain frame. While a single leader whose risk preferences flip-flop to cohere to the popular choice in each frame might benefit from agreeing with most observers *within* each frame, it's reasonable to expect that observers might detect the inconsistency of the leader when confronted with the choices *across* both frames. Prior research makes clear that observers disparage leaders who make inconsistent choices or change their minds, even when they agree with the new stance espoused by the leader (Dorison, Umphres, & Lerner, 2021; Effron, Lucas, & O' Connor, 2015; Effron, O'Connor, Leroy, & Lucas, 2017; Jordan, Sommers, Bloom, & Rand, 2017; Kreps, Laurin, & Merritt, 2017). Observers may therefore no longer penalize leaders for demonstrating consistent risk preferences compared to leaders who demonstrate inconsistent risk preferences. However, this strategy relies on constituents noticing decision frames (and in turn reducing biased risk preferences of constituents)—a possibility that is far from guaranteed (but see Frisch, 1993).

If they do not want to rely on constituents passively noticing their (lack of) susceptibility to framing effects, leaders may simply want to explain framing effects to constituents—thus also reducing biased risk preferences of constituents (along Path A). However, a communication strategy based solely on identifying the underlying decision bias may prove ineffective for multiple reasons. First, the communication strategy may simply have no effect on the tendency to demonstrate the bias, especially for biases that occur outside of conscious awareness (for discussion, see Arkes, 1991; Fischhoff, 1982; Yoon, Scopelliti, & Morewedge, 2021; Larrick, 2004; Stanovich, 1999). For example, in the context of halo effects (i.e., the tendency to use global impressions to generate evaluations of unrelated individual attributes; Nisbett & Wilson, 1977; Landy & Sigall, 1974; Thorndike, 1920), informing and warning participants about the phenomenon has negligible impact on its persistence (Wetzel, Wilson, & Kort, 1981; Wilson & Brekke, 1994; see also Loewenstein, Bryce, Hagmann, & Rajpal, 2015). Second, even

once an observer *knows* about the bias, she still may *perceive* the leader as untrustworthy. For example, Kreps, Laurin, and Merritt (2017) found that observers disliked leaders who changed their moral minds—even when they agreed with the new stance of the leader (for related work, see Zlatev, 2019). Put another way, even when observers agreed with the leader's choice, the choice itself may signal an underlying negative trait about the leader (e.g., that they are cold or lack integrity, despite being competent). For these reasons, an informational explanation simply communicating to observers about a decision bias (such as framing effects) may be ineffective.

Formally, I hypothesized:

Hypothesis 2: Strategies focused on shifting constituents' biased risk preferences will produce minimal benefits.

Instead, I propose that a more effective strategy is one focused on *reducing penalties from disagreement* (i.e., a strategy focused on Path B). While there are many such potential communication strategies, one that I draw from the research literature on conflict resolution is expressing learning goals (Collins, Dorison, Gino, & Minson, 2022; for related work, see Yeomans et al., 2020; Minson, Chen, & Tinsley, 2020; Minson & Chen, 2022; Hussein & Tormala, 2021). Prior research in the context of disagreement over partisan politics in the United States reveals that beliefs that a counterpart holds learning goals decreases derogation and increases willingness to engage in the future. While presently untested, the same might be true for leaders defending their risk preferences to skeptical observers.

Why might this be the case? I theorized that expressing learning goals builds trust by counteracting the negative attributions that typically result from disagreement. That is, it could minimize the costs of disagreement—even while having negligible persuasive impact. If this latter hypothesis is supported, we would expect that expressing learning goals would have minimal effect on the probability that an observer agrees with the choice of the leader. Instead, we would expect that expressing learning goals would have a relatively larger benefit among observers who disagree (vs. agree) with the choice of the leader. Expressing learning goals could make disagreement more palatable, thus allowing leaders to build trust when making unpopular choices (while accruing smaller benefits among those already

supportive). Beyond trust, I also measured willingness to work for the leader in the future, and expected all patterns with trust to replicate with this additional workplace measure.

Formally, I hypothesized:

Hypothesis 3a: Expressing learning goals will increase perceived trustworthiness of a leader. This effect will be stronger among observers who disagree (vs. agree) with the leader's risk decision.

Hypothesis 3b: Expressing learning goals will have negligible impact on the probability that an observer agrees with the leader's risk decision.

Research Overview

Four experiments examine whether leaders can predict and overcome costs for effective decision making under risk. Examining a unique sample of police executives, Experiment 1 investigates whether experienced leaders can forecast the reputational costs of ignoring loss-gain decision frames. Experiments 2-3 test whether attempting to shift constituents' biased risk preferences overcomes these penalties. Finally, Experiment 4 tests whether expressing learning goals can do so by minimizing reputational costs for disagreement, especially for unpopular choices. Across experiments, I explore the central yet complex role of biased risk preferences of both leaders and constituents as a driving mechanism underpinning these effects.

I report how I determined sample size, all exclusions, all manipulations, and all measures in all experiments (Simmons et al., 2012). Preregistrations, materials, data, and code are available here: https://researchbox.org/1936&PEER_REVIEW_passcode=PUSMMQ. Experiments 2 and 4 were pre-registered (Logg & Dorison, 2021).

Experiment 1

Experiment 1 investigates the extent to which leaders anticipate that followers will distrust them for failing to update their choices based on the decision frame. Experiment 1 also assesses whether leaders' forecasts are correlated with their own internal risk preferences. To test these possibilities, I recruited a sample of highly experienced police executives.

Method

Participants. 94 police executives were recruited as part of an executive leadership institute between 2022-2023. Designed for police executives, the course includes a rigorous set of training in a variety of decision making and leadership capabilities. Although the study context did not provide for the capture of specific demographic information, participants were predominantly male and typically had 20-30 years of policing experience. Sample size was determined based on logistics of the course.

Procedure. At the beginning of the executive education course, police executives read the classic Influenza Problem developed by Kahneman and Tversky (1981; described below). Experiment 1 took place during a relative decline in COVID-19.

Police executives were randomly assigned to one of two between-subjects experimental conditions. They read: "Imagine the U.S. is preparing for the outbreak of a *new* strain of the flu, which is expected to kill 600 people in this country. There are two alternative programs." In the Gain Condition, they read: If Program A is adopted, 200 people will be saved. If Program B is adopted, there is a one-third probability that all 600 people will be saved and a two-thirds probability that no people will be saved." In the Loss Condition, they read: "If Program C is adopted, 400 people will die. If Program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die." Based on participant feedback, the executives agreed that these types of high-stakes, risky decisions are similar to those they and their colleagues (e.g., other police chiefs) regularly face during their professional responsibilities.

I collected two sets of outcome variables, both of which were answered as binary response options (i.e., the participant had to select either Program A/C or Program B/D, depending on experimental condition). First, in line with the canonical research on loss-gain framing effects on risk preferences, I asked police executives what they themselves would choose to do. I expected that—replicating a voluminous prior literature—their risk preferences would shift depending on the frame in which the problem was presented, such that executives in the Gain Condition would prefer the risk-averse option (i.e., Program A) and executives in the Loss Condition would prefer the risk-seeking option (i.e., Program D).

Second, and more novel to the present investigation, I asked the police executives to predict which policy option would engender greater trust from the public. Specifically, using a binary response format (i.e., the executives again had to select either Program A/C or Program B/D, depending on experimental condition), I asked which choice they thought would make them look more trustworthy. I theorized that their forecasts regarding perceived trust would mirror their own risk preference—and that they would thus accurately forecast how their choices would be perceived, such that their perceived reputational incentives shift depending on how the options were framed.

Results

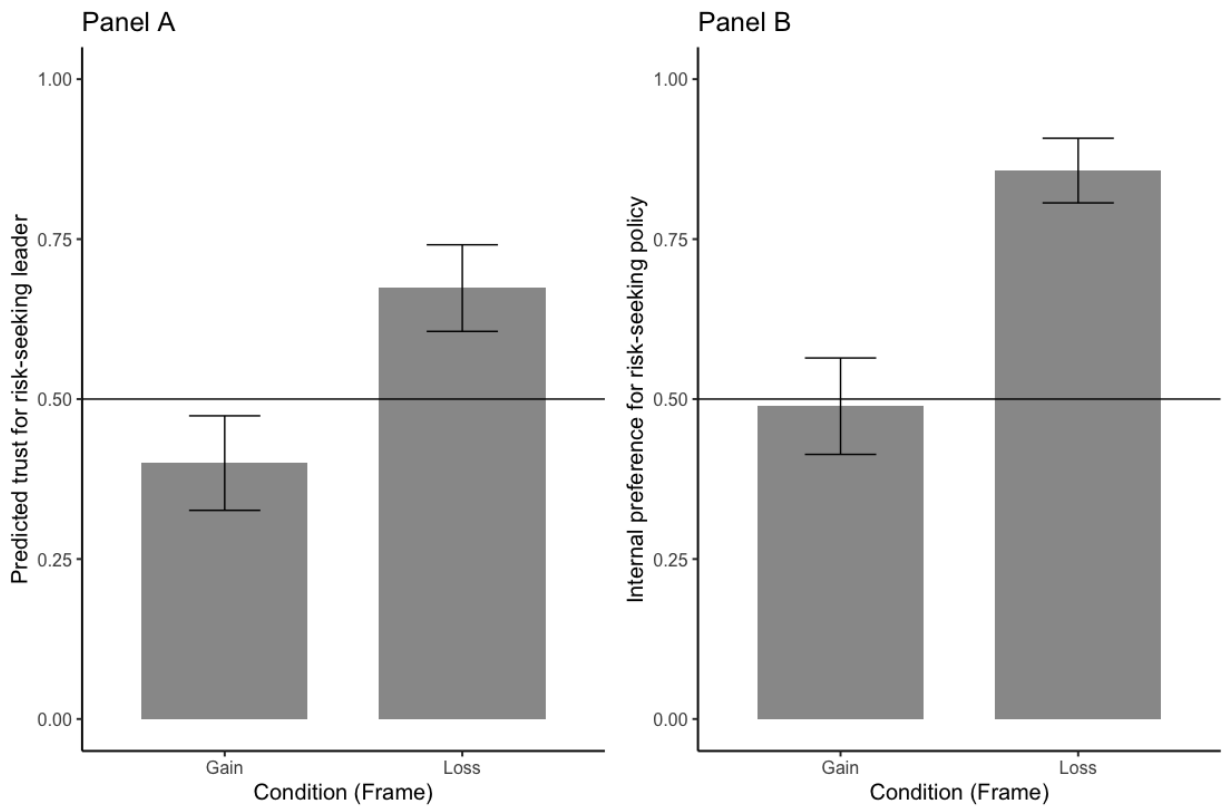
The primary research question of Experiment 1 was whether police executives anticipated that the reputational consequences of their choices would depend on the frame in which options were presented. Supporting Hypothesis 1, and in contrast to prior literature demonstrating systematic failures in how individuals forecast how they will be perceived by others, they anticipated that their reputational incentives shifted based on decision frames. In the Gain Condition, 40% of executives predicted that implementing the risk-seeking policy would make them look more trustworthy. However, in the Loss Condition, 67% of executives predicted that implementing the risk-seeking policy would make them look more trustworthy ($t(92) = 2.74$, $p = .007$, Cohen's $D = 0.56$). Indeed, the frame in which the options were presented *reversed* which policy choice police chiefs predicted would engender trust—in line with the reality of their reputational incentives. Given the results of past research (Dorison & Heller, 2022; see also Supplemental Experiment 1), this reversal implies an underlying accuracy of their predictions. Results are depicted in Figure 3 (Panel A).

I next analyzed to what extent, if at all, experienced police executives fall victim to loss-gain framing effects on their own internal risk preferences. Supporting Hypothesis 1A, and in line with a large prior research literature (Camerer & Hogarth, 1999; Dawes, Faust, & Meehl, 1989; Enke et al., 2023; Northcraft & Neale, 1987; Massey & Thaler, 2013; McNeil, Pauker, Sox, & Tversky, 1982; Schwitzgebel & Cushman, 2015; Shanteau, 1992; Simmons & Massey, 2012), they did: whereas 51% of police executives preferred the risk-seeking program when choice outcomes were framed as gains, 86%

preferred the risk-seeking program when choice outcomes were presented as losses. These were significantly different from each other ($t(92) = 4.12$, $p < .001$, Cohen's $D = 0.85$). As an intriguing aside, police executives in general tended to be slightly risk-seeking (grand mean = 68% chose the risk-seeking option across frames; comparison to chance: $t(93) = 3.74$, $p < .001$); however, they still demonstrated a clear and robust effect of loss-gain framing on risk preferences, as described above. Results are depicted in Figure 3 (Panel B).

Supporting Hypothesis 1B, predicted trust from the public was strongly correlated with internal risk preferences ($r = .47$, $p < .001$). Put another way, police executives' choices of which program to select converged with the program they thought would make them look more trustworthy 73% of the time. This pattern, in which police chiefs' forecasts of their reputational incentives are correlated with their own (biased) risk preferences, supports predictions grounded in past research on bias blind spot (Ehrlinger, Gilovich, & Ross, 2005; Pronin, Lin, & Ross, 2002; Scopelliti et al., 2015) and naïve realism (Ross & Ward, 1995; Schwalbe, Cohen, & Ross, 2016). Indeed, just as the (biased) internal risk preferences served as the foundation for trust judgments of observers in past research (and Supplemental Experiment 1), the same internal (biased) risk preferences served as the foundation for predicted trust among leaders in Experiment 1. As an exploratory analysis, I tested whether leaders' own preferences mediated the effect of condition (loss vs. gain) on predicted trust from the public. This analysis returned a significant indirect effect (beta = .17, z = 3.01, p = .003). I continue to explore this central role for biased risk preferences across experiments.

Figure 3. In Experiment 1, police chiefs predicted that their reputational incentives would reverse depending on whether choices were presented in the gain vs. loss frame (Panel A). Critically, these predictions followed closely with their own (biased) risk preferences (Panel B).



Discussion

Experiment 1 tested whether police executives predict that their reputational incentives shift depending on whether choice outcomes are framed as losses or gains. Using the canonical Influenza Problem developed by Kahneman and Tversky (1979, 1981), results revealed that they do. Of note, their forecasts closely paralleled their own biased risk preferences. In sum, they both fell victim to loss-gain framing effects and (accurately) expected to be trusted for doing so.

Experiment 1 raises the question of how leaders can overcome the predictable costs associated with ignoring decision frames. Leaders' decisions are not typically evaluated in isolation. Instead, they can explain, justify, and defend their choices to observers over time. How should they do so? Experiments 2-4 investigate this question.

Experiment 2

Experiment 1 provided evidence that leaders (accurately) forecast that their reputational incentives shift depending on whether options are presented as losses or gains. However, both Experiment

1 and past research investigate evaluations of a *single* choice made by a leader within a *single* frame. Over time, such dynamics might be different. For example, a key open question remains whether observers reward an *individual leader* whose risk preferences shift depending on decision frames. Put another way, would individuals trust a single leader who "flip flopped" depending on how options are presented, as long as the leaders' choices were congruent with those typically held by observers? Or, would consistent risk preferences allow a leader to maintain trust over time? Is simply being consistent an effective strategy to overcome the reputational costs of maintaining consistent preferences across decision frames?

Method

Participants. I recruited 880 individuals living in the United States from Prolific Academic (mean age = 37.92, age range = 18-85, 48% female) in July 2023. As pre-registered, I included only the 834 participants (95%) who correctly answered the simple attention check before random assignment. Exclusions did not vary by condition and statistical significance of results remains unchanged when analyzing the full sample. The same is true in all experiments. Sample size was determined based on predicted effect sizes from prior research (Dorison & Heller, 2022).

Procedure. After giving informed consent and answering a quick attention check (in which they were instructed to select a specific answer from a list of ten if they were reading the instructions), participants learned that they would be evaluating the decisions made by a hypothetical political candidate: Candidate Williams.

As a test of generalizability, I adapted the canonical Influenza Problem developed by Kahneman and Tversky to a new context: economic policymaking by a public leader. All participants read that twice a year, the Bureau of Labor Statistics releases a new report and policy options. Participants then read that in the first half of 2022, the report predicted that 600,000 people would lose their jobs. The report identified two policy options in response. Participants in the Loss Condition read that if Policy A is implemented, 400,000 people will lose their jobs for sure; alternatively, if Policy B is implemented, there is a 1/3 chance that no one will lose their jobs and a 2/3 chance that 600,000 people will lose their jobs. In

the Gain Condition, participants learned about the same two policies, but with the policies framed in terms of jobs saved. Specifically, participants in the Gain Condition read that if Policy A is implemented, 200,000 people will keep their jobs for sure; alternatively, if Policy B is implemented, there is a 1/3 chance that 600,000 people will keep their jobs and a 2/3 chance that no one will keep their jobs.

Order was counterbalanced such that half of the participants saw the loss frame first and the other half of participants saw the gain frame first. After reading about each policy option, participants learned which policy Candidate Williams endorsed (either the risk-seeking or risk-averse policy). They then indicated how effective of a leader they perceived Candidate Williams to be on a sliding scale from 0-100 and indicated which policy they themselves preferred.[2] This measure of leadership effectiveness is closest conceptually to what prior literature terms ability-based trust (rather than benevolence- or integrity-based trust).

On the following screen, participants learned that in the second half of the same year, the Bureau of Labor Statistics again releases a new report and associated policy options. Critically, the report again predicted that 600,000 people would lose their jobs and again identified two new policy options in response. Policy C was equivalent to Policy A and Policy D was equivalent to Policy B; however, the frame in which the policies were presented was the opposite of whatever the participant saw in the first half of the year (i.e., if policy options in the first half of the year were in the loss frame, then policy options in the second half of the year were in the gain frame, and vice versa). Participants again learned which policy Candidate Williams endorsed (either the risk-seeking or risk-averse policy), again evaluated Candidate Williams on the same sliding scale, and again indicated which policy they themselves preferred.

The primary dependent variable was the average of the two evaluations. In addition, I pre-registered to also assess each rating individually in a secondary analysis. Supplemental Experiment 2 replicates these findings with a measure of perceived trustworthiness.

[2] I mistakenly pre-registered that this item would reference trust rather than an overall evaluation of leadership effectiveness. I measured trust in a replication experiment (Supplemental Experiment 2, described later).

Participants were randomly assigned to one of four between-subjects experimental conditions which varied the policy choices of Candidate Williams. In the "frame-congruent flip-flopper" condition, Candidate Williams made a risk-seeking choice in the loss frame and a risk-averse choice in the gain frame. In the "consistent risk-seeker" condition, Williams made a risk-seeking choice in both frames. In the "consistent risk-avoider" condition, Williams made a risk-averse choice in both frames. Finally, in the "frame-incongruent flip-flopper" condition, Williams made a risk-averse choice in the loss frame and a risk-seeking choice in the gain frame.

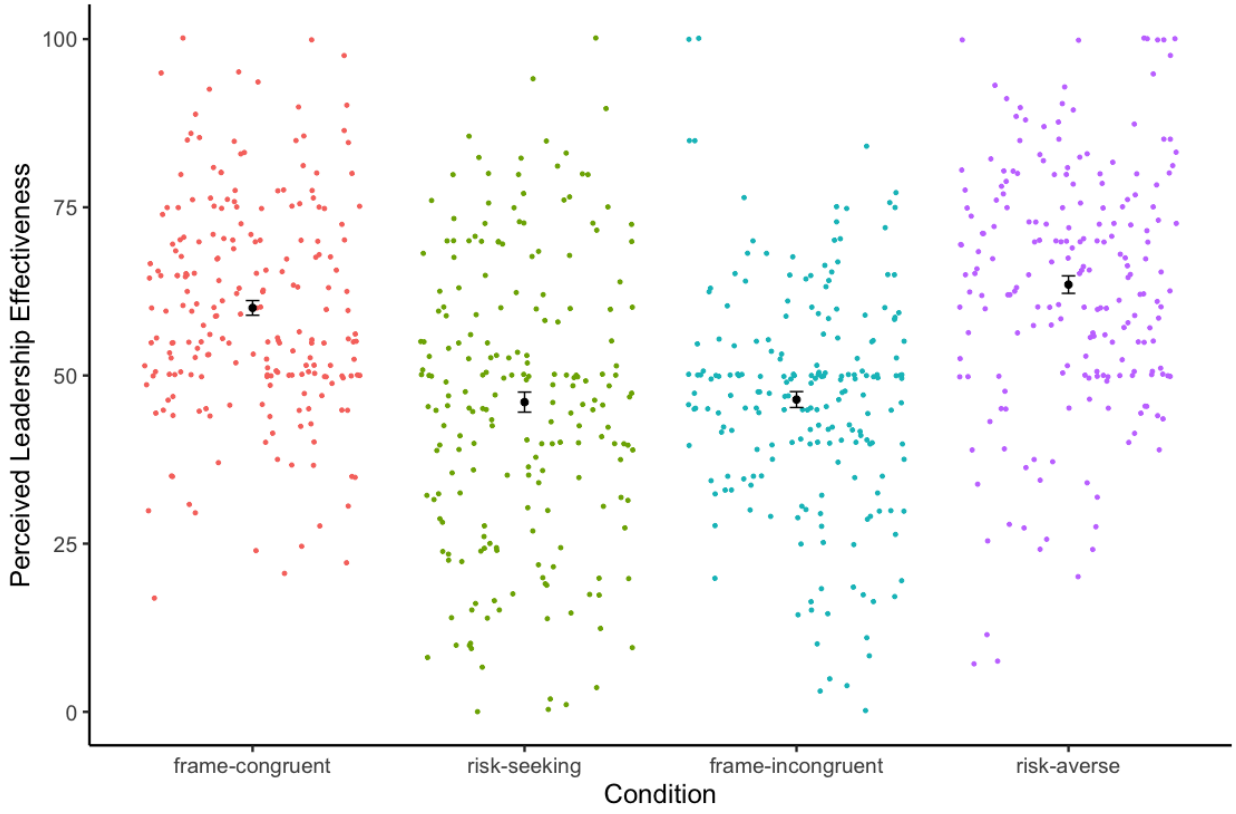
Finally, participants answered a few short demographic questions (i.e., age and gender). Participants had the opportunity to leave open-ended responses about the survey before receiving a code for payment.

Results

Experiment 2 tested Hypothesis 2, examining how observers' evaluations shifted depending on the risk preferences of the leader and as a function of participants' own internal risk preferences.

I regressed overall evaluations (i.e., the average of the two individual ratings) on condition, where the reference group was the frame-congruent flip-flopper. This regression yielded three coefficients, each assessing one of the other conditions against the key reference group. As depicted in Figure 4, results supported the key predictions: the frame-congruent flip-flopper was perceived as a more effective leader than both the consistent risk-seeker ($M_{congruent}$ = 60.05 vs. $M_{seeking}$ = 46.04, t = 7.85, $p < .001$) and the frame-incongruent flip-flopper ($M_{congruent}$ = 60.05 vs. $M_{incongruent}$ = 46.42, t = 7.65, $p < .001$). The frame-congruent flip-flopper was evaluated similarly, although slightly less positively, than the consistent risk-avoider ($M_{congruent}$ = 60.05 vs. M_{averse} = 63.51, t = 1.94, p = .052). In an additional experiment using an identical design but trustworthiness as the outcome variable (Supplemental Experiment 2, N = 802), I again found that frame-congruent leaders were perceived more positively than the frame-incongruent and risk-seeking leaders, but that they were perceived similarly to the risk-averse leader (see *SI* for full details). Thus, frame-congruent flip-flopping held a substantial advantage over consistent risk-seeking, while yielding little-to-no disadvantage against consistent risk-avoidance.

Figure 4. In Experiment 2, observers evaluated the frame-congruent flip-flopper more positively than the consistent risk-seeker and the frame-incongruent flip-flopper. The frame-congruent flip-flopper was evaluated similarly to the consistent risk-avoider. Additional analyses revealed that participants held strong internal preferences for the risk-averse policy option, which underpinned these asymmetric leadership evaluations. Error bars represent 1 SE and colored dots represent raw data.



What explains this pattern of results? In line with the conceptual model developed in the introduction, I hypothesized that the effect of leader choice on evaluations would be underpinned by how frames influence participants' own internal risk preferences, such that participants positively evaluated leaders who make the same choice they themselves preferred. To test this conceptual model, I had three specific predictions and concomitant analyses.

First, based on data from Supplemental Experiment 2 (which was conducted before this experiment), I expected participants to have a strong internal preference for the risk-averse (vs. risk-seeking) option in the gain frame but to be relatively evenly split in the loss frame. Put another way, I

expected more participants to hold frame-congruent internal preferences in the gain frame than in the loss frame, where I defined a frame-congruent preference as risk-aversion in the gain frame and risk-seeking in the loss frame. Participants received a score of 1 if they held a frame-congruent preference and a score of 0 if they held a frame-incongruent preference. Each participant thus received two scores—one per frame. I then conducted a paired-sample t-test where the dependent variable was whether the participant held a frame-congruent preference (1 = yes, 0 = no) and the independent variable was the frame (1 = gain, 0 = loss). Results supported the prediction, such that participants held an internal preference for the frame-congruent choice 80% in the gain frame, but only 50% in the loss frame ($t(833) = 11.59$, $p < .001$). That is, participants showed stronger risk-aversion in the gain frame than they did risk-seeking in the loss frame.

Second, as a direct result of the overall risk-aversion of the sample, I expected participants to agree with the leader most often in the frame-congruent and risk-averse conditions, and relatively less often in the frame-incongruent and risk-seeking conditions. To investigate this possibility, I created a new variable indicating agreement with the leader for each choice. This combined index, which measured total agreement with the leader, had scores of 0 (disagree on both choices), 1 (agree on one choice and disagree on one choice), and 2 (agree on both choices). Using the identical regression set-up to the analyses with evaluations above, I then regressed agreement with the leader on condition, using the frame-congruent leader as the reference group. As predicted, results revealed a negative regression coefficient when assessing the frame-congruent condition against the risk-seeking condition ($M_{congruent} = 1.41$ vs. $M_{seeking} = 0.83$, $t = 9.41$, $p < .001$) and the frame-incongruent condition ($M_{congruent} = 1.41$ vs. $M_{inconguent} = 0.83$, $t = 9.40$, $p < .001$). There was no difference with the risk-averse condition ($M_{congruent} = 1.41$ vs. $M_{averse} = 1.48$, $t = 1.06$, $p = .29$). Mirroring the pattern with global evaluations above, I thus found that frame-congruent flip-flopping held a substantial advantage over consistent risk-seeking and frame-incongruent flip-flopping, while yielding little-to-no disadvantage against consistent risk-avoidance.

Third, I expected agreement with the choice of the leader to be strongly associated with evaluations. Regressing evaluations on agreement indicated this was the case (unstandardized beta =

17.75, t = 23.21, $p < .001$). Put another way, and collapsing across condition, participants gave an average overall leadership evaluation of 69.6 when they agreed with both choices made by the leader, 51.2 when they agreed with one choice made by the leader, and just 34.3 when they agreed with neither choice made by the leader. Clearly, evaluations depended heavily on agreement with the choice of the leader.

Finally, as an exploratory analysis, I also assessed whether such effects held when examining solely the *second* choice of the leader that was evaluated. That is, it could be the case that the benefits for frame-congruent flip-floppers are accrued exclusively when observers evaluate their first choice; when the second choice comes around, it could be that observers recognize the inconsistency of the frame-congruent flip-flopper, and the benefits are attenuated. I did not find evidence to support this hypothesis. Instead, as with the combined evaluations, the frame-congruent flip-flopper was perceived more positively than the consistent risk-seeker ($M_{congruent} = 57.35$ vs. $M_{seeking} = 45.07$, $t = 5.75$, $p < .001$) and the frame-incongruent flip-flopper ($M_{congruent} = 57.35$ vs. $M_{incongruent} = 44.98$, $t = 5.85$, $p < .001$), and slightly more negatively than the consistent risk-avoider ($M_{congruent} = 57.35$ vs. $M_{averse} = 62.27$, $t = 2.31$, $p = .021$). Supplemental Experiment 2 replicated the relative benefit for frame-congruency against risk-seeking and frame-incongruency (all ps $< .001$), but did not replicate the relative cost against risk-aversion. Thus, given the overall convergence between first and second evaluations, leaders did not appear to pay an additional hypocrisy penalty even directly after their inconsistent risk choices were revealed.

Discussion

Experiment 2 provides converging evidence that leaders whose risk preferences are inconsistent (thus violating the statistical axiom of transitivity) are perceived positively, sometimes more so than those whose risk preferences are consistent. The effect of leader choice on leadership evaluations was underpinned by how frames influence participants' own internal risk preferences, such that participants positively evaluated leaders who make the same choice they themselves prefer. While in this sample participants were relatively risk-averse, it is important to note that the results suggest there is nothing inherently reputationally beneficial about risk-averse leaders. Instead, the specific pattern of results is likely to depend on the specific context and risk preferences of the audience. I discuss the central role for

constituent risk preferences across experiments and return to this point in further depth in the General Discussion.

Experiment 3

Experiment 3 investigates whether providing a very short description of framing effects eliminates constituents' preferences for leaders who demonstrate them. I predicted that even after being provided information about framing effects, constituents would still trust a leader more who considers decision frames. Of note, Experiment 3 harnessed a large-scale, nationally representative sample of adults in the United States. By doing so, it provides increased precision in estimating effect sizes (due to the large sample size: over 20 times that used in prior experiments) and increased generalizability (due to the nationally representative nature of the population sampled).

Method

Participants. 24,957 participants were recruited between December 2022 - February 2023 as part of the COVID States Project, a large-scale internet survey conducted by an academic consortium in the United States. 20,477 participants answered the focal dependent variable (described below) and are included in the final dataset. Participants were recruited from all 50 states (and Washington DC), were at least 18 years old, and lived in the United States at the time of taking the survey. Participant demographics were matched to representative quotas on age, gender, race and ethnicity, and geographic distribution. For more information on the participant sample, see here: www.covidstates.org. Sample size was determined by administrators from the COVID States Project.

Procedure. As part of the COVID States Project, participants answered dozens of questions related to COVID attitudes and behavior, political beliefs, and trust in institutions. All questions can be found here: https://www.covidstates.org/tags.

I added one relevant question, which was asked about midway through the survey. Specifically, drawing on canonical descriptions of framing effects (e.g., Bazerman & Moore, 2009; Ruggeri et al., 2021; Tversky & Kahneman, 1981), all participants read the following: "According to psychologists, a framing effect is a tendency for people to decide on options based on whether the options are framed with

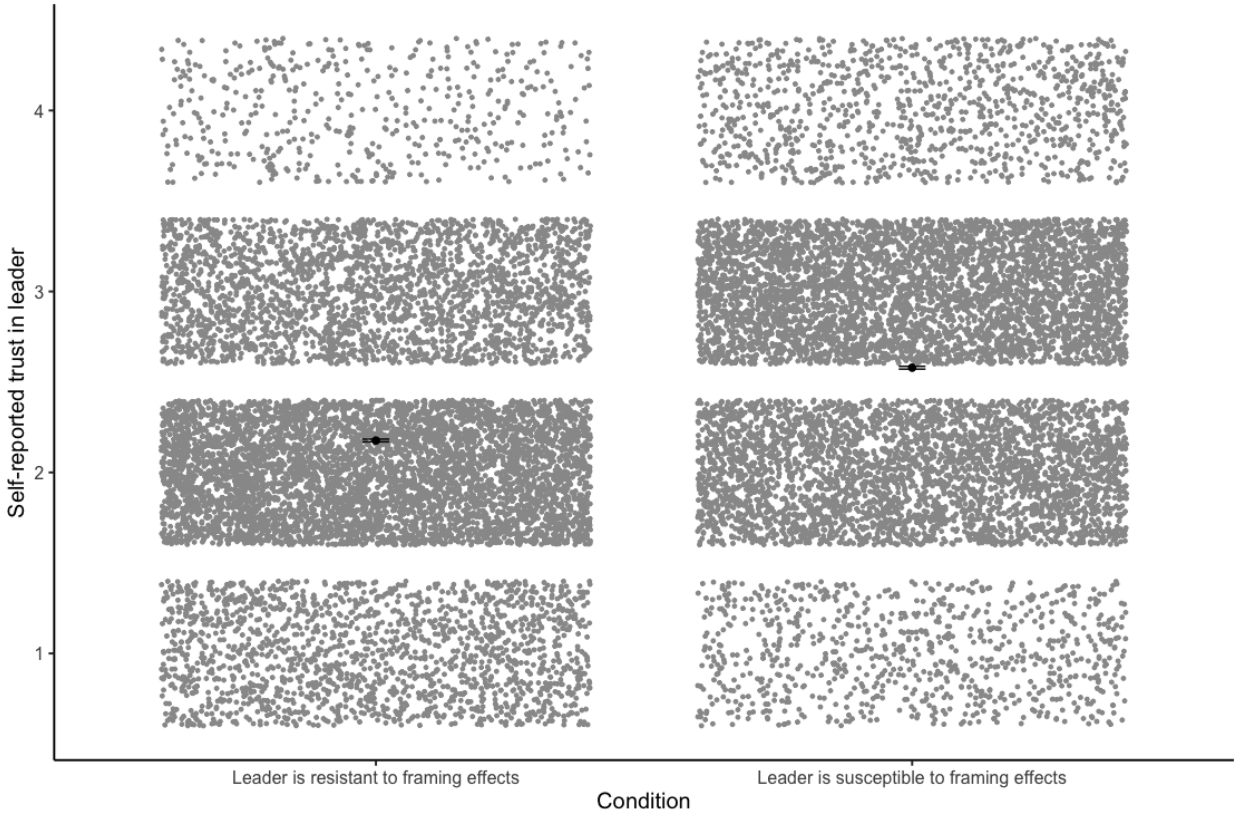
positive or negative connotations; e.g., as a loss or as a gain." Participants were then randomly assigned to one of two between-subjects experimental conditions. In the Susceptible Leader Condition, I asked participants: "How much would you trust a leader who considers how the options are framed when making decisions?" In the Resistant Leader Condition, I asked participants: "How much would you trust a leader who does not consider how the options are framed when making decisions?" Participants answered the question on a 4-point scale with labels of not at all, not too much, some, and a lot. I predicted that participants would give higher ratings to leaders who was susceptible (vs. resistant) to how options are framed, even directly after learning about framing effects (for related work, see Frisch, 1993).

Results

The focal question was whether providing a short description of framing effects eliminates (or even reverses) citizens' preferences for leaders who demonstrate them. Supporting Hypothesis 2, it did not. Instead, participants demonstrated a robust preference for a hypothetical leader who was susceptible (vs. resistant) to how options are framed ($Mean_{Susceptible} = 2.58$ vs. $Mean_{Resistant} = 2.18$, $t(20475) = 37.22$, $p < .001$, Cohen's $D = 0.52$). Results are depicted in Figure 5.

The results were not only statistically significant (a relatively low bar considering the sample size of over 20,000 participants), but also practically significant. To put these results in perspective, I ran a simulation in which I randomly drew 100,000 pairs of participants, one who evaluated a leader in the Susceptible Condition and one who evaluated a leader in the Resistant Condition. For each randomly drawn pair, I then assessed how often the susceptible leader was evaluated more positively than the resistant leader. Participants who evaluated a susceptible leader gave higher trust ratings 48% of the time, and the reverse just 20% of the time (the remaining 32% of pairs indicated equal levels of trust). Put another way—and temporarily ignoring the cases in which the pairs indicated equal levels of trust—participants who evaluated the susceptible leader were twice as likely to rate the leader as more trustworthy than those who evaluated the resistant leader (48% vs. 20%, respectively).

Figure 5. In a large and nationally representative sample (Experiment 3), participants who were informed about framing effects trusted a leader who was susceptible to them more than one who was resistant to them. Error bars represent 1 SE and gray dots represent raw data.



Discussion

Harnessing data from a large-scale nationally representative survey, Experiment 3 provides evidence that simply informing citizens about framing effects does not eliminate a preference for leaders who are susceptible to them. Instead, participants perceived a hypothetical leader who was susceptible to framing effects as more trustworthy than one was resistant to them—and this effect was meaningful in magnitude. Building off Experiment 2, Experiment 3 thus provides converging evidence that the reputational costs of ignoring decision frames are relatively sticky (either in the form of joint evaluation, as was the case in Experiment 2, or learning about them directly before evaluating a hypothetical leader, as was the case in Experiment 3). Experiment 4 investigates a theoretically driven means of avoiding such reputational costs.

An alternative explanation for the pattern of results in Experiment 3 is that observers simply prefer their leaders to consider multiple factors. Experiment 4 also addresses this possibility.

Experiment 4

Experiments 1-3 and Supplemental Experiments 1-2 revealed that (1) observers systematically penalize leaders whose risk preferences are unaffected by loss-gain framing, (2) leaders predict such penalties, and (3) simply being consistent or informing observers about framing effects does not serve to mitigate such penalties. How then should leaders navigate this tension between building trust and following value-maximizing decision processes? Leaders need tools to communicate their value-maximizing decisions in a way that maintains trust. Experiment 4 draws on literature on conflict resolution to test such a tool: expressing learning goals (Collins, Dorison, Gino, & Minson, 2022; Hussein & Tormala, 2021). Experiment 4 also continues to investigate a central role for constituents' internal risk preferences. Finally, to address the alternative explanation from Experiment 3 described above, I also tested an additional communication strategy focused on considering multiple factors.

Three additional methodological revisions merit note. First, to assess generalizability, I again adapted the classic framing scenarios to a new context: a private-sector CEO of a healthcare technology company. Second, I measured not only trust, but also willingness to work for the CEO's future entrepreneurial venture. Finally, to examine ecological validity (while acknowledging the limitations inherent in controlled laboratory experiments), I recruited both a general population sample and a sample selected for entrepreneurial work experience.

Method

Participants. I recruited 2167 individuals living in the United States (mean age = 43.44, age range = 18-94, 46.35% female) in July 2023 through Prolific Academic. Experiment 4 contains a sub-sample recruited from the general participant population and a sub-sample limited to those with prior or current entrepreneurial experience. I found no meaningful differences between the two sub-samples, so collapse across them below (online materials give full details). As pre-registered, I included only the 2033 participants (94%) who correctly answered the simple attention check before random assignment. Sample

size was determined to power a relatively small interaction, although I did not have an estimate of the precise effect size when conducting this experiment.

Procedure. The procedure of Experiment 4 builds on that of Experiments 1-2, with a few key design revisions. After giving informed consent and answering a quick attention check (in which they were instructed to select a specific answer from a list of ten if they were reading the instructions), participants learned that they would read a hypothetical scenario regarding two potential policies that a CEO was deciding between.

Participants were randomly assigned to one of three primary between-subjects experimental conditions. The conditions varied the communication strategy used by the CEO (control vs. learning goals vs. consider factors). In the control condition, participants simply saw a choice made by the CEO; no justification was given. In the learning goals condition, in addition to seeing the choice, participants read the following text communicated by the CEO: "I understand that other people might make the opposite choice and I would be curious to hear about why. I think it would be interesting to understand that perspective better." In the consider additional factors condition, the CEO gave the following justification: "I considered many dimensions when making my choice, including how the decision was framed. I think my choice is sound and will benefit the organization."

The vignette in Experiment 4 was designed to adapt the canonical decision making under risk paradigm to an organizational context. All participants read about the CEO of a technology company, Hal Arizin, who was debating between two different investments for the company. While I was primarily interested in assessing the impact of communication strategy in general, I also varied two additional factors: the frame in which the choice was made (loss vs. gain) and the choice made by the CEO (risk-seeking vs. risk-averse). I considered both experimental manipulations secondary in this experiment because I was primarily interested in the effect of communication strategy on trust.

The vignette proceeded as follows. In all conditions, participants read the following: Hal Arizin is the CEO of a large healthcare company named Acme Technology. Mr. Arizin faces a difficult decision, where he must choose between two different and mutually exclusive investment options. In the gain

condition, participants read: "If he selects Investment A, then Acme will make 2 million dollars for sure. If he selects Investment B, there is a one third probability that Acme will make 6 million dollars and a two thirds probability that Acme will make $0 (i.e., break-even)." In the loss condition, participants read a similar paragraph, but with outcomes framed in terms of losses: "If he selects Investment A, then Acme will lose 4 million dollars for sure. If he selects Investment B, there is a one third probability that Acme will lose $0 (i.e., break-even) and a two thirds probability that Acme will lose 6 million dollars.[3]

I measured two primary dependent variables. First, I asked participants how much they trusted Mr. Arizin on a 0-100 slider. In addition, I was also interested in whether such effects would persist to a downstream organizational outcome: willingness to join a start-up that Mr. Arizin would be launching soon. Specifically, I asked participants (approximately half of whom had prior or current entrepreneurial experience) how interested they would be in joining a future company of Mr. Arizin's. This question was answered on the same 0-100 slider. I expected all patterns with trust to replicate with this workplace measure, and for the two measures to be strongly correlated.

I collected one additional variable: which investment choice participants themselves preferred. This measure allowed me to test competing hypotheses regarding the mechanism through which expressing learning goals could operate, detailed below.

Finally, participants answered a few short demographic questions (i.e., age and gender). Participants had the opportunity to leave open-ended responses about the survey before receiving a code for payment.

Results

The primary question in Experiment 4 was when and why expressing learning goals benefits leaders when making decisions under risk. I conducted three sets of analyses to answer this overarching question.

[3] An important methodological difference with prior studies is that the present vignette constitutes a reflection effect (rather than a pure framing effect) because the gambles differ in the sign of the outcome itself (i.e., money lost vs. gained compared to break-even), rather than representing isomorphic outcomes framed differently (Fagley, 1993).

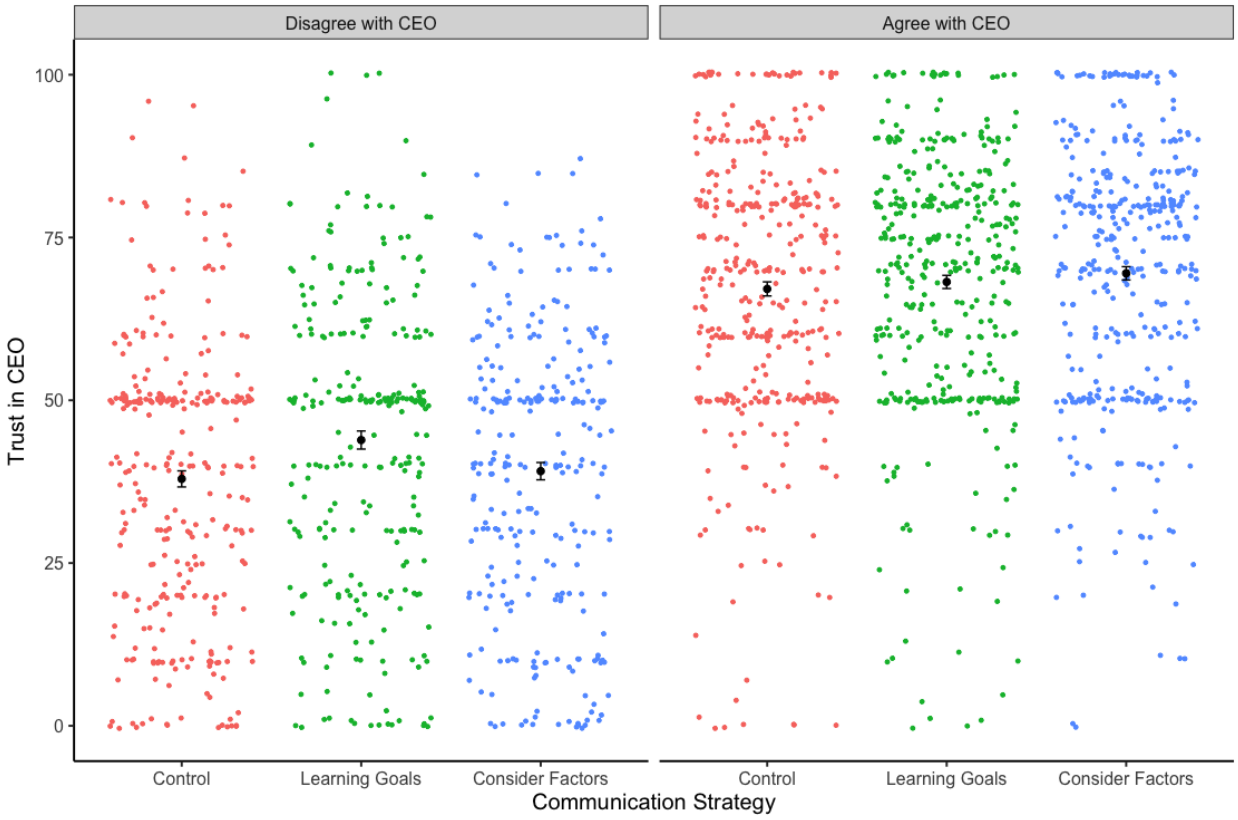
The first set of analyses tests for an overall effect of expressing learning goals. To do so, I regressed trust on condition, where the reference group was the control condition. This regression yielded two coefficients, each assessing one of the other conditions against the reference group. Results revealed that both expressing learning goals ($M_{control}$ = 54.3 vs. $M_{LearningGoals}$ = 58.4, p = .0029) and considering additional factors ($M_{control}$ = 54.3 vs. $M_{ConsiderFactors}$ = 57.1, p = .043) increased trust compared to the control condition, although the effect of expressing learning goals was approximately 47% larger (mean difference = 4.1 vs. 2.8). Expressing learning goals significantly increased willingness to work for the CEO's future venture, as well ($M_{control}$ = 49.2 vs. $M_{LearningGoals}$ = 53.5, p = .0032). Considering additional factors had a directional, but not significant, benefit ($M_{control}$ = 49.2 vs. $M_{LearningGoals}$ = 51.0, p = .21). Trust and willingness to join the CEO's future venture were strongly correlated (r = .83, p < .001).[4]

The second and third set of analyses examine *when and why* expressing learning goals fosters trust. Building on Experiments 1-2 (and Supplemental Experiments 1-2), I continue to assess a central role for agreement between an observer's internal risk preferences and the leader's choice, here as a function of communication strategy employed by the leader. More concretely, the second set of analyses examined whether expressing learning goals has persuasive effects (i.e., decreases the amount of disagreement). To test this possibility, I regressed agreement on condition, where the reference group was again the control condition. In line with predictions, I found negligible effects of either expressing learning goals ($M_{Control}$ = 56.1% vs. $M_{LearningGoals}$ = 59.7%, p = .18) or considering additional factors ($M_{Control}$ = 56.1% vs. $M_{ConsiderFactors}$ = 59.0%, p = .27) on agreement with the choice of the leader. As an intriguing aside, overall agreement across conditions was above chance levels ($M_{Overall}$ = 58.2%; comparison to chance: p < .001), suggesting some overall social learning effects of observing the leader's choice (c.f., Yoon, Scopelliti, & Morewedge, 2021).

[4] Exploratory follow-up analyses using the same regression set-up but with learning goals as the hold-out reference group revealed that expressing learning goals slightly, although certainly not robustly, outperformed considering additional factors for both trust (p = .043) and future work intentions (p = .088)

Finally, the third set of analyses examines whether expressing learning goals provides relatively greater benefits when the observer disagrees (vs. agrees) with the choice of the leader. Results are depicted in Figure 6. To do so, I regressed trust on condition, agreement with the leader, and their interaction. Compared to the control condition, when the observer disagreed with the choice of the CEO, the CEO expressing learning goals increased both trust ($M_{control}$ = 37.9 vs. $M_{LearningGoals}$ = 43.9, p = .0013) and willingness to work for the CEO's future venture ($M_{control}$ = 32.8 vs. $M_{LearningGoals}$ = 38.1, p = .0078). There was no such effect when the observer agreed with the choice of the CEO for either trust ($M_{control}$ = 67.1 vs. $M_{LearningGoals}$ = 68.2, p = .47) or future work intentions ($M_{control}$ = 62.0 vs. $M_{LearningGoals}$ = 64.0, p = .22). This resulted in a significant interaction for trust (beta = 4.90, p = .036) and a directional, but not significant, interaction for future work intentions (beta = 3.36, p = .19). I found no similar asymmetry for considering additional factors, which not only did not produce a significant benefit for either trust or future work intentions at either level of agreement, but also did not produce a significant interaction for either variable (all ps > .10).

Figure 6. In Experiment 4, expressing learning goals (but not considering additional factors) increased trust from observers, especially observers who disagreed with the risk choice of the CEO. Error bars represent 1 SE and colored dots represent raw data.



Discussion

Experiment 4 demonstrates a successful communication strategy leaders can use to overcome the reputational costs associated with having consistent risk preferences. Expressing learning goals increased trust and willingness to work for a CEO's future venture, especially among constituents who disagreed with the leader's risk choice. Of note, such effects did not rely on minimizing disagreement; rather, they worked by reducing the costs of disagreement. Further, they did not generalize to a different communication strategy (considering additional factors).

General Discussion

Leaders must navigate risk and uncertainty to create value for their organizations and society. And yet, these aims can systematically conflict with the goal of building trust. Managing the reputational penalties associated with high-quality decision processes represents a major challenge for leaders. Do leaders predict these penalties? What communication strategies can they use to overcome them? And what are the key underlying mechanisms at play?

Four experiments address these questions in the widely influential context of loss-gain framing effects on risk preferences (Tversky & Kahneman, 1981). First, Experiment 1 reveals both that leaders can forecast the reputational costs of ignoring loss-gain frames on risk preferences and that their forecasts of the reputational consequences of their choices were correlated with their own (biased) risk preferences. Second, Experiment 2 demonstrates that ignoring decision frames yields reputational costs even when a focal leader maintains consistency across frames. These penalties were especially harsh for risk-seeking leaders and were underpinned by observers' own (biased) risk preferences. Finally, while Experiment 3 revealed that simply informing observers about framing effects had negligible impact, Experiment 4 successfully tested a strategy leaders can use to maintain trust, especially when making unpopular risk choices: communicating learning goals. As a complement to individual-level training, the present research suggests a necessary focus on identifying organizational norms and developing communication strategies to help leaders manage and overcome canonical decision biases.

Theoretical contributions

The present research integrates research from behavioral science, communication, and leadership to provide novel insights for managerial decision making. In doing so, it contributes to the literature in several ways. First, the present work contributes to a robust and interdisciplinary research literature on decision making under risk (Caraco, Martindale, & Whittam, 1980; Gigerenzer et al., 1999; Loewenstein, Webber, Hsee, & Welch, 2001; Mishra, 2014; Tversky & Kahneman, 1979). Understanding how individuals navigate trade-offs among prospects has yielded critical theoretical insights for managerial judgment and decision making. Traditionally, research on decision making under risk strips away the social context (for important exceptions, see Anderson & Galinsky, 2006; Van Kleef et al., 2021). Following this theoretical approach, prior influential research has shed enormous light on the underlying cognitive processes that yield systematic errors and biases that violate traditional statistical axioms (Kahneman & Tversky, 1979). Building on this important work, the present research embeds decision making under risk in a broader social and organizational context. The findings highlight the importance of

understanding the role that social and reputational factors can play in driving managerial decision making (Tetlock, 2000).

Second, the present work contributes to behavioral science research on leadership. Past research has found that while it is possible to debias choices, it is often difficult to do so (Morewedge et al., 2015; Sellier, Scopelliti, & Morewedge, 2019; see also Chang, Chen, Mellers, & Tetlock, 2016; Fong, Krantz, & Nisbett, 1986; Fong & Nisbett, 1991; Ho, Budescu, Dhami, & Mandel, 2015; Larrick, Morgan, & Nisbett, 1990). The present research identifies a key underlying role for such biased risk preferences: they not only underpin trust judgments made by observers, but also underpin predicted trust among leaders.

Developing communication strategies that shift the reputational incentives associated with different choices is a novel socially situated approach. It may be particularly promising for leaders, who rely on trust from followers to lead successfully and maintain desired outcomes of power and status (Galinsky & Schweitzer, 2016). Critically, the present work builds on the foundation of past research that examined the myriad ways that accountability structures in organizations can shape judgment and choice (Lerner & Tetlock, 1999; Tetlock, 2000). The present research highlights the need for future research examining when leaders can accurately predict their reputational incentives, and to what extent these predictions drive their judgments and decisions.

Third, leaders not only make decisions that steer their organizations, but also justify and explain their choices to key constituencies. How can reputational penalties for unpopular risk decisions be alleviated? The present work provides a nuanced answer. On the one hand, the conclusions are somewhat pessimistic: simple and straightforward informational communication strategies appear ineffective. Yet, this is perhaps unsurprising: many decision biases occur outside of conscious awareness (Arkes, 1991) and warning individuals about cognitive biases often has negligible impact on their own tendency to display them (e.g., Wetzel, Wilson, & Kort, 1981; Wilson & Brekke, 1994; Loewenstein, Bryce, Hagmann, & Rajpal, 2015). Indeed, even if an observer agrees with the leader's choice, they may still evaluate them negatively, insofar as the choice signals other negative underlying traits about the leader, such as hypocrisy (e.g., Kreps, Laurin, & Merritt, 2017).

On the other hand, the conclusions are somewhat optimistic. Expressing learning goals allowed leaders to maintain trust by making disagreement more palatable (i.e., the reduced the reputational costs of disagreement), without relying on persuasive impact (i.e., without reducing the amount of disagreement). Such communication strategies can overcome the key leadership challenge associated with following value-maximizing decision processes while maintaining trust from key constituents. Aligning such incentives is a critical tool for leaders.

Limitations and future directions

The present research has several limitations, which in turn serve as the foundation for future directions. First, I primarily investigated how observers form immediate impressions of a leader at the time a risk-seeking (or risk-averse) choice is made. Effects outside of a laboratory context are likely to be smaller than those documented here when observers have more information about the leader and/or a pre-existing history, as is often the case in organizations. However, leaders care not only about internal stakeholders (where they often have extensive pre-existing history), but also about external stakeholders—many of whom are forming first impressions. Individuals often form impressions of leaders who they have never met or evaluated first-hand, in a different frame than that encountered by the leader, or over the span of days or weeks. Future work could examine how leaders' reputations spread across social networks in the form of gossip (for related work, see Craik, 2009; Costello & Srivastava, 2021), how individuals shift decision frames (for related work, see Daniels & Zlatev, 2019; Zlatev, Daniels, Kim, & Neale, 2019), and how such impressions morph over time. All of these are important empirical questions that the present research asks but does not answer.

Second, while the present research included multiple measurements of trust, it did not parse sub-facets of trust within a single study. Prior research by Levine and colleagues makes clear that certain behaviors, such as prosocial lies, can have opposing influences on different sub-facets of trust (Levine & Schweitzer, 2015; see also Berman, Levine, Barasch, & Small, 2015; Huppert et al., 2023). How might such patterns play out in the context of decision making under risk? Future work could draw on the research literature on halo effects to generate novel hypotheses. Specifically, this research literature posits

that individuals use global impressions to generate evaluations of individual attributes. In perhaps the most famous demonstration, Nisbett and Wilson (1977) tested whether evaluations of a college professor's warm (vs. cold) manner could in turn shape evaluations of unrelated attributes of that professor (e.g., physical attractiveness). They did. Such effects are robust across contexts and persist even when individuals are warned about them (Wetzel, Wilson, & Kort, 1981). In the context of decision making under risk, it could be the case that penalties for disagreement would be quite broad in nature, influencing multiple sub-facets of trust approximately equally. But future empirical work is needed to directly test this possibility.

Third, while the present research harnessed multiple representative samples of individuals in the United States, such dynamics might vary across audiences or cultures. In the present studies, observers tended to be risk-averse, underpinning a general preference for risk-averse leaders (c.f., Van Kleef et al., 2021). In contrast, employees in an organization with an extensive training program in probability and statistics or in an organization with a highly risk-seeking (or risk-averse) culture might show an attenuation of such preferences, or even a reversal. In these cases, the frame in which choice options are presented would likely explain less variance in observer evaluations. However, even in cases in which internal employees are relatively immune to framing effects, leaders must still often consider how they will be evaluated by external stakeholders (e.g., investors or the public). Future research is needed to assess these boundary conditions and tease apart how leaders weigh these different constituencies.

Fourth, while I identified the expression of learning goals as a strategy leaders can use to overcome the reputational costs of having consistent risk preferences, this is certainly not the only effective strategy. Indeed, given the relatively modest effect sizes found in Experiment 4, other strategies (or combinations of strategies) might be even more effective. Future research should examine which communication tools leaders are already using and how they can be sharpened not only through expressing learning goals, but also through other communication strategies. The present work suggests that communication strategies focused on minimizing the trust penalty for disagreement, rather than the amount of disagreement, hold promise.

Finally, a notable methodological limitation with the present work is its exclusive reliance on laboratory experiments. While I recruited multiple representative samples of laypeople, used relatively large financial stakes (compared to the related research literature), and recruited a unique sample of experienced police executives, lab experiments still yield constraints on ecological and external validity. Future research is needed, for example, to determine which communication strategies leaders naturally use to explain and justify their risk preferences. Building on this idea, follow-up work could examine the text of speeches on the congressional floor from lawmakers, earnings calls from CEOs, or tweets from presidential candidates and code whether they express learning goals or other related receptive cues.

Conclusion

From the founder of a technology start-up to the president of a nation, leaders must navigate decisions under conditions of risk and uncertainty to maximize value—while simultaneously maintaining trust from key constituencies. The present research not only finds that these goals systematically conflict (and that leaders can forecast this conflict), but also identifies and develops an effective communication strategy leaders can use to resolve the conflict. Broadly, the present research contributes to a growing body of research linking behavioral science, communication, and leadership (Moore & Bazerman, 2022).

References

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*, 718–735.

Anderson, C., & Galinsky, A. D. (2006). Power, optimism, and risk-taking. *European Journal of Social Psychology*, *36*(4), 511-536.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, *110*, 486–498.

Arrow, K. J. (1974). *The limits of organization*. WW Norton & Company.

Bazerman, M. H. (1983). Negotiator judgment: A critical look at the rationality assumption. *American Behavioral Scientist*, *27*(2), 211-228.

Bazerman, M. H., & Moore, D. A. (2013). Judgment in Managerial Decision Making. John Wiley & Sons.

Benartzi, S., & Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *The Quarterly Journal of Economics*, *110*(1), 73-92.

Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The Braggart's Dilemma: On the Social Rewards and Penalties of Advertising Prosocial Behavior. *Journal of Marketing Research*, *52*(1), 90–104

Blunden, H., Logg, J. M., Brooks, A. W., John, L. K., & Gino, F. (2019). Seeker beware: The interpersonal costs of ignoring advice. *Organizational Behavior and Human Decision Processes*, *150*, 83–100.

Brockner, J. (1992). The escalation of commitment to a failing course of action: Toward theoretical progress. *Academy of Management Review*, *17*(1), 39-61.

Camerer, C. F., & Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, *19*(1), 7–42.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2017). Statistically inaccurate and morally unfair judgements via base rate intrusion. *Nature Human Behaviour*, *1*(10).

Caraco, T., Martindale, S., & Whittam, T. S. (1980). An empirical demonstration of risk-sensitive foraging preferences. *Animal Behaviour*, *28*(3), 820-830.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, *11*(5), 509–526.

Chen, M. K., Lakshminarayanan, V., & Santos, L. R. (2006). How basic are behavioral biases? Evidence from capuchin monkey trading behavior. *Journal of Political Economy*, *114*(3), 517-537.

Collins, H. K., Dorison, C. A., Gino, F., & Minson, J. A. (2022). Underestimating Counterparts'

Learning Goals Impairs Conflictual Conversations. *Psychological Science*, *33*(10), 1732-1752.

Costello, C. K., & Srivastava, S. (2021). Perceiving personality through the grapevine: A network approach to reputations. *Journal of Personality and Social Psychology*, *121*, 151–167.

Craik, K. H. (2008). *Reputation: A Network Interpretation*. Oxford University Press.

Daniels, D. P., & Zlatev, J. J. (2019). Choice architects reveal a bias toward positivity and certainty. *Organizational Behavior and Human Decision Processes*, *151*, 132–149.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, *243*(4899), 1668–1674.

Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: meta-analytic findings and implications for research and practice. *Journal of Applied Psychology*, *87*(4), 611.

Dorison, C. A., & Heller, B. H. (2022). Observers penalize decision makers whose risk preferences are unaffected by loss–gain framing. *Journal of Experimental Psychology: General*, *151*, 2043–2059.

Dorison, C. A., Umphres, C. K., & Lerner, J. S. (2022). Staying the course: Decision makers who escalate commitment are trusted and trustworthy. *Journal of Experimental Psychology: General*, *151*, 960–965.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*, 380–417.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Effron, D. A., Lucas, B. J., & O'Connor, K. (2015). Hypocrisy by association: When organizational membership increases condemnation for wrongdoing. *Organizational Behavior and Human Decision Processes*, *130*, 147-159.

Effron, D. A., O'Connor, K., Leroy, H., & Lucas, B. J. (2018). From inconsistency to hypocrisy: When does "saying one thing but doing another" invite condemnation? *Research in*

Organizational Behavior, *38*, 61-75.

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin*, *31*(5), 680-692.

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., & Van De Ven, J. (2023). Cognitive biases: Mistakes or missing stakes?. *The Review of Economics and Statistics*, *105*(4), 818-832.

Epley, N., Kardas, M., Zhao, X., Atir, S., & Schroeder, J. (2022). Undersociality: Miscalibrated social cognition can inhibit social connection. *Trends in Cognitive Sciences*, *26*(5), 406-418.

Epley, N., & Schroeder, J. (2014). Mistakenly seeking solitude. *Journal of Experimental Psychology: General*, *143*(5), 1980.

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*, 772–787.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slavic and A. Tversky (Eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge University Press.

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes*, 54(3), 399-429.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*(3), 253–292.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, *120*, 34–45.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, *56*(4), 279-304.

Friedman, M., & Savage, L. J. (1952). The expected-utility hypothesis and the measurability of utility. *Journal of Political Economy*, *60*(6), 463-474.

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. *Group Processes & Intergroup Relations*, *8*(2), 109-124.

Galinsky, A., & Schweitzer, M. (2015). *Friend & foe: When to cooperate, when to compete, and how to succeed at both*. Currency.

Gigerenzer, G., Todd, P. M. & the ABC Research Group (1999) *Simple heuristics that make us smart.* Oxford University Press.

Gilovich, T. D., & Griffin, D. W. (2010). Judgment and decision making. In *Handbook of social psychology, Vol. 1, 5th ed* (pp. 542–588). John Wiley & Sons, Inc.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*, *78*(2), 211.

Hershey, J. C., & Schoemaker, P. J. (1980). Risk taking and problem context in the domain of losses: An expected utility analysis. *Journal of Risk and Insurance*, 111-132.

Ho, E. H., Budescu, D. V., Dhami, M. K., & Mandel, D. R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy*, *1*(2), 43–55.

Huppert, E., Herzog, N., Landy, J. F., & Levine, E. (2023). On being honest about dishonesty: The social costs of taking nuanced (but realistic) moral stances. *Journal of Personality and Social Psychology*.

Hussein, M. A., & Tormala, Z. L. (2021). Undermining your case to enhance your impact: A framework for understanding the effects of acts of receptiveness in persuasion. *Personality and Social Psychology Review*, *25*(3), 229-250.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological science*, *28*(3), 356-368.

Jung, M. H., Sun, C., & Nelson, L. D. (2018). People can recognize, learn, and apply default effects in social influence. *Proceedings of the National Academy of Sciences*, *115*(35), E8105-E8106.

Kahneman, D., & Tversky, A. (1979). Prospect theory: A decision making under risk. *Econometrica*, *47*(2), 263–291.

Kanodia, C., Bushman, R., & Dickhaut, J. (1989). Escalation Errors and the Sunk Cost Effect: An Explanation Based on Reputation and Information Asymmetries. *Journal of Accounting Research*, *27*(1), 59–77.

Kardas, M., Schroeder, J., & O'Brien, E. (2021). Keep talking:(Mis) understanding the hedonic trajectory of conversation. *Journal of Personality and Social Psychology*.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*(1), 569-598.

Kreps, T. A., Laurin, K., & Merritt, A. C. (2017). Hypocritical flip-flop, or courageous evolution? When leaders change their moral minds. *Journal of Personality and Social Psychology*, *113*, 730–752.

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology*, *29*(3), 299.

Larrick, R. P. (2004). Debiasing. *Blackwell handbook of judgment and decision making*, 316-338.

Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the Use of Cost-Benefit Reasoning in Everyday Life. *Psychological Science*, *1*(6), 362–370.

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, *81*(1), 146.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological*

Bulletin, *125*, 255–275.

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88-106.

Loewenstein, G., Bryce, C., Hagmann, D., & Rajpal, S. (2015). Warning: You are about to be nudged. *Behavioral Science & Policy*, *1*(1), 35-42.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*, 267–286.

Logg, J. M., & Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes*, *167*, 18-27.

Magee, J. C., & Galinsky, A. D. (2008). Social hierarchy: The self-reinforcing nature of power and status. *The Academy of Management Annals*, *2*(1), 351-398.

Massey, C., & Thaler, R. H. (2013). The Loser's Curse: Decision Making and Market Efficiency in the National Football League Draft. *Management Science*, *59*(7), 1479–1495.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734.

McKenzie, C. R., Leong, L. M., & Sher, S. (2021). Default sensitivity in attempts at social influence. *Psychonomic Bulletin & Review*, *28*, 695-702.

McNeil, B. J., Pauker, S. G., Sox, H. C., & Tversky, A. (1982). On the Elicitation of Preferences for Alternative Therapies. *The New England Journal of Medicine*, *306*(21), 1259–1262.

Mercer, J. (2005). Prospect theory and political science. *Annual Review of Political Science*, *8*, 1-21.

Minson, J. A., & Chen, F. S. (2022). Receptiveness to Opposing Views: Conceptualization and Integrative Review. *Personality and Social Psychology Review*, *26*(2), 93–111.

Minson, J. A., Chen, F. S., & Tinsley, C. H. (2019). Why Won't You Listen to Me? Measuring Receptiveness to Opposing Views. *Management Science*, *66*(7), 3069–3094.

Mishra, S. (2014). Decision-making under risk: Integrating perspectives from biology,

economics, and psychology. *Personality and Social Psychology Review*, *18*(3), 280-307.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions: Improved Decision Making With a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140.

Nisbett, R. E., & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*, 250–256.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes*, *39*(1), 84–97.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369-381.

Roberts, A. R., Levine, E. E., & Sezer, O. (2021). Hiding success. *Journal of Personality and Social Psychology*, *120*(5), 1261.

Ross, L., & Ward, A. (1995). Psychological Barriers to Dispute Resolution. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 27, pp. 255–304). Academic Press.

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M., Gibson, S. P., Jarke, H., Karakasheva, R., Khorrami, P. R., Kveder, J., Andersen, T. L., Lofthus, I. S., McGill, L., Nieto, A. E., … Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, *4*(6).

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley.

Schwalbe, M. C., Cohen, G. L., & Ross, L. D. (2020). The objectivity illusion and voter polarization in the 2016 presidential election. *Proceedings of the National Academy of*

Sciences, *117*(35), 21218-21229.

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, *141*, 127–137.

Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S., & Kassam, K. S. (2015). Bias Blind Spot: Structure, Measurement, and Consequences. *Management Science*, *61*(10), 2468–2486.

Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing Training Improves Decision Making in the Field. *Psychological Science*, *30*(9), 1371–1379.

Sezer, O. (2022). Impression (mis) management: When what you say is not what they hear. *Current Opinion in Psychology*, *44*, 31-37.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, *53*(2), 252-266.

Simmons, J. P., & Massey, C. (2012). Is optimism real?. *Journal of Experimental Psychology: General*, *141*(4), 630.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Available at SSRN 2160588*.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research, 177(3), 1333-1352.*

Stanovich, K. E. (1999). Who is rational?: Studies of individual differences in reasoning. *Psychology Press.*

Steinmetz, J., Sezer, O., & Sedikides, C. (2017). Impression mismanagement: People as inept self-presenters. *Social and Personality Psychology Compass,* 11(6), e12321.

Sun, Y., & Mellers, B. (2016). Trade-upgrade framing effects: Trades are losses, but upgrades are improvements. *Judgment & Decision Making*, *11*(6).

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of

confidence. *Journal of Personality and Social Psychology*, *116*, 396–415.

Tetlock, P. E. (2000). Cognitive Biases and Organizational Correctives: Do Both Disease and Cure Depend on the Politics of the Beholder? *Administrative Science Quarterly*, *45*(2), 293–326. https://doi.org/10.2307/2667073

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25-29.

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4481), 453–458.

Van Kleef, G. A., Heerdink, M. W., Cheshin, A., Stamkou, E., Wanders, F., Koning, L. F., ... & Georgeac, O. A. (2021). No guts, no glory? How risk-taking shapes dominance, prestige, and leadership endorsement. *Journal of Applied Psychology*, *106*(11), 1673.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton university press.

Wetzel, C. G., Wilson, T. D., & Kort, J. (1981). The halo effect revisited: Forewarned is not forearmed. *Journal of Experimental Social Psychology*, *17*(4), 427–439.

White, M., Levine, E., & Kristal, A. (2023). Rules are (often) meant to be broken: The effects of discretion and consistent rule-following on interpersonal trust. https://psyarxiv.com/n8m26/

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117.

Yeomans, M., Minson, J., Collins, H., Chen, F., & Gino, F. (2020). Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, *160*, 131–148.

Yoon, H., Scopelliti, I., & Morewedge, C. K. (2021). Decision making can be improved through observational learning. *Organizational Behavior and Human Decision Processes*, *162*, 155-188.

Zlatev, J. J. (2019). I may not agree with you, but I trust you: Caring about social issues signals integrity. *Psychological Science*, *30*(6), 880-892.

Zlatev, J. J., Daniels, D. P., Kim, H., & Neale, M. A. (2017). Default neglect in attempts at social influence. *Proceedings of the National Academy of Sciences*, *114*(52), 13643–13648.