Mitigating Gig and Remote Worker Misconduct: Evidence from a Real Effort Experiment

Vanessa Burbano, Columbia Business School*

&

Bennett Chiles, Columbia Business School*

Forthcoming, Organization Science, 2021

Employee misconduct is costly to organizations and has the potential to be even more common in gig and remote work contexts, where workers are physically distant from their employers. There is thus a need for scholars to better understand what employers can do to mitigate misconduct in these non-traditional work environments, particularly as the prevalence of such work environments is increasing. We combine an agency perspective with a behavioral relationship-based perspective to consider two avenues through which gig employers can potentially mitigate misconduct: 1) through the communication of organizational values and 2) through the credible threat of monitoring. We implement a real effort experiment in a gig work context that enables us to cleanly observe misconduct. Consistent with our theory, we present causal evidence that communication of organizational values, both externally-facing in the form of social/environmental responsibility and internally-facing in the form of an employee ethics code, decreases misconduct. This effect, however, is largely negated when workers are informed that they are being monitored. We provide suggestive evidence that this crowding-out is due to a decrease in perceived trust that results from the threat of monitoring. Our results have important theoretical implications for research on employee misconduct, and shed light on the tradeoffs associated with various potential policy solutions.

**Keywords:** unethical behavior, employee misconduct, employee cheating, employee governance, organizational values, corporate social responsibility, ethics code, monitoring

*Equal authorship. Please send correspondence to vanessa.burbano@gsb.columbia.edu.
INTRODUCTION

Jobs in which workers are physically distant from their employers are becoming increasingly prevalent. This is in part due to a surge in the “gig” or “sharing” economy\(^1\) (Hagiu and Wright, 2019; Kokkodis and Ipeirotis, 2015; Sundararajan, 2016) and in part due to increases in remote work more broadly. According to the McKinsey Global Institute, roughly 11% of U.S. and European workers earn their fulltime income through gig work, and another 14% participate in the gig economy through part-time or occasional freelance work.\(^2\) Remote work more generally increased 159% in the U.S. between 2005 and 2017 and is expected to become a prominent feature of a “new normal” after covid-19.\(^3\) Both gig and remote work contexts are marked by the fact that workers are physically distant from their employers, resulting in unique employee governance challenges. In particular, this physical separation often makes employee motivation and oversight more difficult (Mann and Holdsworth, 2003; Shamir and Salomon, 1985; Wiesenfeld et al., 1999, 2001; Kurland and Egan, 1999).

Drawing from agency theory (e.g., Ross, 1973; Jensen and Meckling, 1976; Harris and Raviv, 1979; Holmstrom 1979), we posit that employee misconduct is likely to be more prevalent in gig and remote work settings. Specifically, we argue that the physical separation inherent in these contexts intensifies the principal-agent problem in two important ways. First, gig and remote workers are likely to feel less connected to their employing organizations, thus exacerbating the misalignment of interests between workers and the firm. Secondly, physical separation typically comes with a greater degree of information asymmetry. Under these conditions, employee misconduct – already a widespread problem costing U.S. firms as much $600 billion annually in traditional work settings (Murphy, 1993; Shulman, 2012; List and Momeni, 2020; Pierce, Snow and McAfee, 2015) – is likely to be a critical and growing challenge.

\(^1\) The system in which intermediary platform firms connect requesters (e.g., employers or consumers) with on-demand gig workers.
We combine an agency perspective with a relationship-based behavioral perspective to frame our consideration of potential policy solutions in gig and remote work contexts. We thus respond to a call from management scholars to combine agency theory with complementary theories in order to incorporate a more complex and realistic view of human and organizational behavior than that afforded by agency theory alone (Eisenhardt 1989). Specifically, we draw on tenets of agency theory to describe the high-level levers that an employer can pull to mitigate gig and remote worker misconduct: 1) by increasing the interest alignment of goals between the worker and the organization, and 2) by decreasing the degree of information asymmetry between the worker and the organization. While the effects of pecuniary incentives on such levers have been explored extensively (e.g., Frank and Obloj 2014; Gubler et al. 2015; Larkin, 2014; Larkin and Pierce, 2015; Obloj and Sengul 2012; Oyer 1998), there has been relatively less consideration of how relationship-based motivators and incentives can influence these levers and interact with other policies designed to reduce worker misconduct (Flammer and Luo, 2008).

We first propose an understudied avenue through which employers can potentially mitigate gig and remote worker misbehavior: through communication of organizational values, which we argue increases the alignment of interests between worker and employer by fostering a sense of shared values with the firm. Next, since monitoring gig and remote workers is often challenging in practice, we consider how the credibility of the threat of monitoring is likely to affect misconduct in these settings. Finally, we also predict that policies emphasizing organizational values are likely to be less effective when monitoring (or the threat thereof) is also utilized. Specifically, we argue that the threat of monitoring is likely to lower perceived trust between worker and employer. This reduction in trust inhibits workers from forming the sense of shared values with the employer that the communication of organizational values would otherwise elicit – thus reducing the effectiveness of this approach in aligning workers’ interests with those of the organization. This crowding out effect is likely to apply not only to gig or remote work settings, but to traditional work settings as well, and to our knowledge it has not been theoretically examined or empirically tested in either.

Assessing the effectiveness of employer-level policies intended to reduce misconduct is empirically challenging for two critical reasons. First, misconduct is exceedingly difficult to measure with accuracy in
real work settings since workers have the incentive to conceal such behavior (Burbano and Ostler, 2020). Second, employer-level policies such as those outlined above are rarely exogenous in practice – making it difficult to separate their causal impact on misconduct from other potential confounding factors. To sidestep these challenges, we designed a novel real effort experiment in a gig work context which enabled us to observe and accurately measure misconduct, as well as to randomly assign communications regarding employer-level policies.

In our study, workers were hired on an online gig market platform to complete a short-term assignment which entailed entering information into the “Contact” sections of various websites. They were also given the optional opportunity to earn bonus payments by contacting the website owners by phone, with instructions to either leave a scripted voicemail or to obtain answers to a market research survey. Unbeknownst to the workers, we owned and operated both the websites to which they were directed and the corresponding phone numbers listed. As such, we were able to cleanly observe whether workers actually entered the requested information into the website as directed (or shirked on the job), as well as whether or not they fraudulently claimed any bonus payments from their employer.

Workers were randomly assigned to one of six conditions in a three-by-two design. First, workers randomly received a message about the employer’s ethical values (an expression of internally-oriented organizational values), a message about the employer’s social/environmental values (an expression of externally-oriented organizational values), or no additional messaging. Second, we manipulated whether or not workers were informed about the potential that their work would be monitored. Notably, from the workers’ perspective, the credibility of the monitoring threat varied between the primary task (where the threat was not particularly credible) and the bonus task (where the threat was at least partially credible) – allowing us to shed light on the extent to which credibility matters.

We find that when implemented individually, both communication of organizational values and the threat of monitoring are effective at reducing employee misconduct by statistically and economically significant margins. (Indeed, even when not credible, the threat of monitoring reduces misconduct – though the effect is less robust here vs. when the threat is credible.) When implemented in combination, however,
the effects of values-oriented policies and monitoring policies are not additive. Specifically, when the threat of monitoring is in effect, the communication of organizational values does very little to further reduce misconduct. We explore potential underlying mechanisms, and find suggestive evidence consistent with the theory that monitoring erodes the trust needed to establish a sense of shared values between workers and the firm. Indeed, empirical results indicate that the threat of monitoring substantially reduces workers’ perceptions of employer trust. And the communication of ethical and social/environmental values increases workers’ perceptions of shared values with the firm – but only when monitoring is not also in effect.

We extend scholarship on misconduct by combining two perspectives – an agency perspective and a relationship-based behavioral perspective – to develop theory on how employer-level policies are likely to influence gig and remote workers’ organization-harming behavior. In so doing, we apply a more complex view of agent (in our context, gig or remote worker) behavior than that afforded by agency theory alone (Eisenhardt 1989), and identify ways that policies designed to address worker misconduct could backfire when agents behave based on social and relationship-based considerations in addition to economic ones. Moreover, our paper contributes revealed, rather than stated, behavioral evidence on the way in which employer-level characteristics influence misconduct in a real work context – an important contribution to the broader literature on misconduct, which has been limited to date in that it has largely been based on laboratory experiments or self-reported survey data (Edelman and Larkin, 2014; Pierce and Balasubramanian, 2015; Pierce and Snyder, 2008) rather than on behavioral evidence where real work effort and outcomes are observed (Burbano and Ostler, 2020; List and Momeni, 2020). While our paper is focused on gig and remote work contexts, we also discuss the extent to which our results may generalize to more traditional work contexts in our conclusions, arguing that the integration of a relationship-based perspective with agency theory leads to a deeper understanding of the drivers and mitigators of employee misconduct in organizations more broadly.
MITIGATING EMPLOYEE MISCONDUCT

Employees ostensibly engage in misconduct because such behavior benefits them in some way. Workers who shirk, for example, gain time for leisure (or for other work for which they may receive additional compensation), and workers who engage in theft or fraud gain financially, at least if their actions go undetected. Yet while workers may benefit from engaging in misconduct (Hirsh, Lu and Galinsky, 2018), the organizations for which they work are often harmed as a result. While some prior literature has highlighted specific types of employee misconduct that might directly or indirectly benefit the organization (e.g., Bennett, Pierce, Snyder, and Toffel, 2013; Burbano and Ostler, 2020; Pierce and Snyder, 2015; Pinto et al., 2008; Umphress et al., 2010; Vardi and Wiener, 1996), we focus on forms of misconduct that are detrimental to firm performance. From a strategic human capital management perspective, this organization-harming misconduct is arguably the type that firms most seek to curtail.

Organization-harming misconduct occurs in work contexts where the interests of the employing organization and the worker are not naturally aligned. When this misalignment of goals is combined with information asymmetry between the employer and employee, a canonical agency problem arises (e.g., Ross, 1973; Jensen and Meckling, 1976; Harris and Raviv, 1979; Holmstrom, 1979), such that it is often optimal for agents (workers) to make choices that are not in the best interests of the principal (the employing organization). 4

This general framing of employee misconduct is one of the oldest applications of agency theory, and it holds in both traditional workplaces and in gig work environments – though the extent to which information asymmetry and incentive misalignment persists or is the more challenging problem likely differs across various organizational contexts. Our focus in this paper is on gig and remote work, and an important defining characteristic of these work environments is the presence of physical distance between employer and worker. This physical distance exacerbates the principal-agent problem in two important

---

4 In this framing, an employing organization, the principal, engages agents (workers) to act on its behalf. The agents’ precise actions, however, are not observable to the principal in full, at least not without effort or cost on the principal’s part.
ways. First, physically distant workers are likely to feel less connected to, and thus identify less strongly with, their employers compared to traditional workers (Wiesenfeld et al., 1999). This lower level of identification with the organization implies that the interests of workers will be inherently less aligned with those of the firm – increasing the incentives for misconduct. Indeed, in related literature, perception of greater physical distance has been associated with lower worker productivity (Cramton and Webber, 2005), lower willingness to do extra work (Burbano, 2019), lower engagement (Kahn, 1990; Zhuang and Gadiraju, 2019), and lower overall performance (Cramton, 2001; Kanawattanachai and Yoo, 2007). In addition, an obvious consequence of physical distance in many instances is that information asymmetries are amplified. As such, misconduct is likely more rampant in many gig and remote work contexts.

Agency theory implies that organizations can work to mitigate employee misconduct in two key ways: 1) by increasing the extent to which employees’ interests are aligned with those of the firm and/or 2) by reducing information asymmetry. To understand how organizational policies might achieve each of these goals and in turn influence worker behavior, it is important to incorporate a relationship-based behavioral perspective to reflect complexities in human behavior which often depend not only on economic dynamics but also social and relationship-based ones. Indeed, to allow for a more complex view of individual and organizational dynamics, scholars have applied tenets of agency theory in combination with other perspectives to derive implications for organization and strategy theory. Examples include Eisenhardt (1988), who combined agency and institutional theories; Anderson (1985), who coupled agency and transaction costs theories; Eccles (1985) who combined agency with equity theory; and Flammer and Luo (1998) who, most closely related to our paper, draw on tenets of agency theory and introduce the importance of considering relationship-based incentives and motivators in influencing employee adverse behavior.

In what follows, we focus on two potential ways in which gig and remote employers might achieve these aims and thereby mitigate worker misconduct. Importantly, by conceptualizing the decisions of agents as influenced not only by economic considerations but also by social and relationship-based considerations (Dunning et al., 212), we identify new ways that these policy approaches might negatively interact.
Communication of Values as a Tool for Mitigating Employee Misconduct

One way in which organizations can mitigate the agency problem is through the implementation of policies that increase alignment between the goals of workers and those of the firm. In practice, this might be achieved in a number of ways. Various forms of financial incentive structures have been explored extensively (e.g., Frank and Obloj 2014; Gubler et al. 2015; Larkin, 2014; Larkin and Pierce, 2015; Obloj and Sengul 2012; Oyer 1998). While these solutions may be effective by some metrics, such as increased productivity and less shirking, they often come with their own set of challenges, which include the crowding out of intrinsic motivation (Deci et al., 1999; Benabou and Tirole, 2006), the proliferation of gaming behavior (Larkin, 2014), increased fraudulent reporting and dishonesty (Cyranoski, 2006; Balasubramanian et al., 2017), increased employee theft (Chen and Sandino, 2012), and co-worker sabotage (Charness et al., 2013; Flory et al., 2016). Apart from pecuniary incentives, employers may be able to better align employees’ interests with those of the firm using relationship-based motivators and incentives (Flammer and Luo, 2017). Exploration of the non-pecuniary drivers of misconduct suggests that relationship-based motivators and incentives are indeed important. For example, perceptions of unfairness and inequity (Larkin, Pierce and Gino, 2012; Gino and Pierce 2009, 2010a, 2010b; Greenberg 1990, 1993) have been shown to influence individuals’ propensity to shirk, cheat and misreport, for example, due to a tendency to justify dishonest acts based on benefits to others (Wiltermuth 20011; Erat and Gneezy 2012; Wiltermuth, Bennett and Pierce 2013). Likewise, relationship-oriented factors such as a sense of rivalry (Kilduff and Galinsky, 2017; Kilduff et al., 2015), competition (Pierce et al., 2013) and feelings of power relative to others (Dubois, Rucker and Galinsky 2015) have been shown to increase unethical behavior by individuals.

Employers might attempt to better align employees’ interests with those of the firm through the adoption of policies that foster a sense of shared purpose between workers and the organization. One important way in which employers might elicit a sense of shared values is by emphasizing organizational values that workers are likely to share. Examples of this in practice might include communications about a commitment to particular values within the organization (e.g., expectations that employees adhere to ethical
values such as honesty, fairness, etc.), or, alternatively, communications about the firm’s commitment to certain values outside the organization (e.g., through various social and/or environmental initiatives).

To the extent that workers share the expressed values, emphasizing them in communications should increase workers’ perceived value congruence and person-organization fit with their employer. Value congruence has been shown to be correlated with organization-benefiting perceptions (see Arthur et al., 2006 for a meta-analysis), identification with the firm (Tajfel and Turner, 1979) and organizational citizenship behavior (Cable and DeRue, 2002), while value incongruence has been shown to result in organization-harming behavior such as reduced motivation and work quality in gig and short-term work contexts (Burbano, 2020). Importantly for its potential to mitigate misconduct, value congruence and person-organization fit has also been linked with employees’ sense of organizational commitment and support for their employing organization’s objective (Valentine, Godkin and Lucero, 2002; Amos and Weathington, 2008; Kristoff-Brown et al, 2005). By increasing support for the employing organization’s goals and objectives, increased perceived value congruence on the part of the employee with the employer thus increases the alignment of interests between principal and agent. Put plainly, the more that a worker feels a sense of shared values with her organization, the more the employee will support the organization’s interests and goals, and the less likely she will be to engage in behavior that harms that organization through misconduct. We thus predict that communications emphasizing organizational values will mitigate adverse behavior among workers.

**H1**: Communications from the employer that emphasize organizational values will reduce gig/remote worker misconduct.

---

5 While Flammer and Luo (2017) make a compelling case that firms appear to be attempting to use CSR as a governance mechanism to reduce employee misconduct, their data does not allow them to test whether CSR is actually effective in serving this purpose.
This hypothesis is contrary to a related prediction by List and Momeni (2020), who posit that communication about CSR, which by our argument can be used to elicit a sense of shared values between the employer and employee and thus decrease misconduct, should in fact increase worker misconduct by eliciting moral licensing. As we discuss in more detail in the conclusion, our theorizing highlights the importance that an organizational policy such as CSR be communicated in a way that elicits a sense of shared values between the worker and employer. Without achieving this sense of shared values, communication of organizational values will not reduce the misalignment of interests between the worker and firm, which is the key mechanism through which communication of organizational values is likely to decrease misconduct.

In traditional workplaces, espousing organizational values might be one part of a larger set of policies intended to build and reinforce a sense of shared values (Stoner, 1989) and a broader and deeper approach to fostering shared values is likely to be more effective at mitigating misconduct than values-oriented messaging alone. Indeed, it is unclear whether claims, without the backing of actions, should be sufficient in influencing stakeholder behavior (Bromley and Powell, 2012). In the case of gig and remote work, where building a strong sense of organizational identification is challenging due to the physical separation of workers from the employer and from one another (Wiesenfeld, Raghuram, and Garud, 1999; Bartel, Wrzesniewski and Weisenfeld, 2012; Petriglieri, Ashford, and Wrzeniewski, 2019), understanding the potential effectiveness of values-oriented messaging as a standalone policy tool is particularly important. Moreover, to the extent that there are measurable effects from messaging alone, such effects can be viewed as a lower bound for the potential benefit when a more comprehensive set of policies is available to be deployed in concert.

**Threat of Monitoring as a Tool for Mitigating Employee Misconduct**

In addition to policies which increase the degree of interest alignment between workers (agents) and the firm (the principal), another mechanism through which firms can mitigate misconduct is by decreasing the degree of information asymmetry between workers and the organization. Monitoring has
long been a commonly proposed solution for reducing adverse behavior by this means (e.g., Becker, 1968). When the principal can observe what the agent is actually doing, this curbs agent opportunism because the agent realizes that she cannot deceive the principal (Eisenhardt, 1989). Essentially, increased monitoring of employee behavior by organizations directly reduces information asymmetry between principal and agent. Under these circumstances, workers should thus engage in less misconduct, since the extent to which they are able to misbehave without detection and subsequent consequences is reduced. Indeed, monitoring has been shown to be effective at reducing employee misconduct in a range of empirical settings (Hubbard, 2000; Nagin et al., 2002; Detert et al., 2007; Olken, 2007; DeHoratius and Raman, 2008; Duflo et al., 2012; Pierce et al., 2015).

Actual monitoring, however, requires the ability to directly observe individuals’ actions – an undertaking that is simply not feasible in some work contexts. In gig and remote work settings where the principal and agent do not share a physical space or interact in person, for example, there is notably less opportunity for an employer to observe the agents’ behavior than in settings where workers and employers share a physical space and interact frequently in person. Even within gig and remote work contexts, there is variance in the feasibility of monitoring worker behavior across settings. In some physically distant worker arrangements, technology can enable observation akin to that which would be achieved by sharing a common physical space. For example, Uber can monitor drivers’ driving patterns, time spent on the road, etc. In such worker arrangements, monitoring is more easily implemented. In many others, though, it is difficult for (current) technology to facilitate the same degree of observation that can be conducted in shared physical space and in contexts of constant interaction between the employee and employer. It is challenging for Upwork and Freelancer.com to accurately ascertain the amount of productive time spent working on jobs contracted through their online platforms, for example. Moreover, even in cases when direct monitoring is feasible, it is often costly for firms to implement (e.g., Dickens et al. (1989) note that many firms expend substantial resources on employee monitoring).

In gig or remote work contexts where actual monitoring is often infeasible or may be too costly to implement, it is possible that the mere threat of monitoring could achieve similar goals. The threat of
monitoring could heighten workers’ sense of organizational oversight and thus decrease workers’ perceptions of informational asymmetry between themselves and their employer (even if the actual amount of informational asymmetry remains the same). This would increase the workers’ perceptions of the expected probability of being caught and punished for misconduct, which is likely to deter them from engaging in misconduct. The threat of monitoring is only likely to achieve these goals, however, if is perceived to be credible. If not perceived to be credible, the threat of monitoring would arguably neither affect workers’ perception of the amount of information asymmetry that exists between them and their employer, nor affect their expectations about the likelihood that they would be caught. Thus, if gig or remote workers are discerning and perceive that the threat of monitoring is not credible given the physically distant nature of the work, we might expect no effect at all on misconduct. The expected effect of the threat of monitoring on gig or remote worker misconduct, then, depends on the credibility of the threat:

\[ H2: \text{The threat of monitoring will reduce gig/remote worker misconduct when the threat of monitoring is credible, but not when the threat of monitoring is not credible.} \]

Finally, even when direct monitoring is feasible and not prohibitively costly to implement or the threat of monitoring is feasible and likely to be perceived to be credible, it is important to note that (the threat of) monitoring may be accompanied by unwanted negative consequences. Bernstein (2012), for example, finds monitoring to be associated with increased levels of gaming behavior among workers, as well as reduced problem-solving capabilities. Frey (1993) and Litzky et al. (2006) argue that monitoring destroys trust between workers and the firm, decreasing intrinsic motivation and organizational commitment. Others have found that monitoring results in increased levels of employee stress and lower job satisfaction (e.g., Aiello and Svec, 1993; Tabak and Smith, 2005). The potential for these negative side effects must be taken into account as employers consider whether or not a monitoring policy is preferable to alternative approaches and/or compatible with other policies that the firm may be simultaneously pursuing. In particular, when conceptualizing the behavior of agents as influenced by social and
relationship-based dynamics in addition to economic dynamics (Dunning et al, 2012), (the threat of) monitoring may carry unintended consequences that weaken the mechanisms through which other simultaneous policy efforts operate, and we address one particular instance of this possibility next.

**The Interaction between (the Threat of) Monitoring and the Communication of Values**

We have argued that both the communication of organizational values and the threat of monitoring should each independently reduce worker misconduct. Given that each is directed at a distinct driver of misconduct identified by agency theory (the threat of monitoring directed at reducing perceived information asymmetry, and the communication of organizational values directed at reducing the misalignment of interests between the principal and agent), it seems reasonable that firms seeking to reduce misconduct might be most effective by implementing policies which address each of the drivers. We thus now focus on how effective these policies might be at reducing misconduct when used in combination with one another.

Of particular concern is the possibility that the threat of monitoring will signal to workers that the organization does not trust them (Frey, 1993; Cialdini, 1996; Litzky et al. 2006; Ferrin et al., 2007). Indeed, control systems more generally have been shown to result in lower levels of trust and, subsequently, less voluntary cooperative behavior (e.g., Malhotra and Murnighan, 2002; Mulder et al., 2006). Communication of organizational values, when fostering a sense of shared values, improves interest alignment between the worker and employer, and intrinsically motivates workers to voluntarily want to behave better towards their employer. (The threat of) monitoring, on the other hand, reduces misconduct because it decreases the (perception of) informational asymmetry between the worker and employer, which reduces misconduct because it makes workers feel as though they have to behave better towards their employer, or else risk being caught and punished. While this may induce "forced” good behavior on the one hand, it is likely that it will simultaneously erode the likelihood of additional “voluntary” good behavior. That is, when a control system that destroys trust like (the threat of) monitoring is utilized, the sense of shared values, and subsequent increase in alignment of goals between the principal and agent, is eroded, such that workers have no intrinsic desire to behave better. This argument is consistent with work which has shown trust to
help explain the effect of value congruence between the employee and organization (Edwards and Cable, 2009), as well as to facilitate or hinder effects of other determinants on worker outcomes (Dirks and Ferrin, 2001; Brockner and Siegel, 1996).

In summary, (the threat of) monitoring destroys trust between workers and the firm. We posit that this lack of trust, in turn, inhibits workers from forming the sense of shared values and interest alignment with the firm that the communication of organizational values would otherwise elicit. Thus, we expect that monitoring will negatively moderate the benefits that organizations might gain from the communication of values with regard to worker misconduct.

**H3: The communication of organizational values will be less effective at reducing gig/remote worker misconduct when the threat of monitoring is also in effect.**

While our focus in this paper is primarily on physically distant gig or remote work, this reasoning should apply to traditional work contexts as well. Furthermore, given that destruction of trust as a result of monitoring is exacerbated in contexts where the relationship between the principal and agent is more interpersonal (Frey, 1993), it likely that this undesirable moderating effect would be even greater in traditional, physically close, work contexts. To our knowledge, this hypothesis has not been theorized nor empirically tested in either traditional or physically-distant work contexts.

**EXPERIMENTAL SETTING AND DESIGN**

Our study was conducted in May 2019 on Amazon Mechanical Turk (MTurk), a large, prototypical crowd-based gig work platform (Scholz, 2017). Like most major players in the gig economy, MTurk is an intermediary platform that connects requesters (gig employers) with on-demand gig workers who complete jobs that are task-based and short-term, while physically distant from their employer (Kaine and Josserand, 2016).

---

6 IRB approval was obtained. The experiment was also pre-registered on Open Science Framework under the project “The Effect of CSR, Ethics Codes, and Monitoring on Employee Misconduct,” available at osf.io/2q5cw.
2019; Meijerink and Keegan, 2019). It facilitates the use of human intelligence to perform tasks for requesters (Kaine and Josserand, 2019); as such, MTurk jobs are commonly referred to as HITs (human intelligence tasks). Typical jobs include simple data entry and survey completion tasks.

The experiment that we implement follows a three-by-two design, as illustrated in Figure 1. We operationalize communication of organizational values in two ways: 1) communication of ethical values, an expression of mainly internally-facing organizational values, and 2) communication of social/environmental values, an expression of mainly externally-facing organizational values. This design allows us to observe the main effects of communication of organizational values (H1) and the threat of monitoring (H2), as well as any potential interaction effects between the communication of organizational values and the threat of monitoring (H3).

*** Insert Figure 1 Here ***

Figure 2 details the basic design and logistics of our study. Acting as a hiring employer (specifically, a start-up firm offering a software product for interior designers), we advertised a data entry job estimated to take 5-10 minutes to complete. To help ensure that workers were unaware that they were participating in a research study, we used the identity of a real start-up organization when hiring workers (with the organization’s approval). Importantly, during the timeframe that the study was run, this company’s website indicated that it was “under construction,” preventing workers from discovering any potential confounding information about the hiring employer online. Payment for the job was $1.50, with additional opportunities to earn bonus payments. The pay and nature of the job were designed to be typical of jobs in the MTurk context.

Workers interested in the job could click on the link included in our MTurk posting to receive further instructions on an external site before deciding whether or not to proceed. Neither the initial MTurk posting nor the initial instructions on our external site contained any language associated with the treatment.

---

7 The actual median time that workers took to complete the job was roughly 10 minutes.
8 Given that all workers were presented with the same employing organization and website, any effects of the website or name of the organization on workers would be constant across conditions.
conditions; rather, treatment occurred only after workers began the job so as to mitigate potential selection effects. Only after choosing to continue the task were workers randomly assigned to one of our six conditions.

*** Insert Figure 2 Here ***

After random assignment, workers received subsequent information about the employer and the job corresponding to their assigned condition. Those in the ethical values (EV) conditions were informed that the hiring company believes in a culture of accountability and transparency in the workplace, and strives to be honest, ethical, and fair. Those in the social/environmental values (SEV) condition were informed that the hiring company believes in giving back to the community and in improving the environment, donating 5% of its profits to that end. Wording in both the EV and SEV conditions was constructed to be representative of how companies typically describe such initiatives. For the SEV condition in particular, we used language corresponding to a common type of externally-facing CSR initiative, namely charitable giving to benefit the community and environment. Finally, those in the threat of monitoring conditions were informed that 5% of HITs would be randomly selected to be reviewed for data quality. See Figure 2 for the exact wording used in each condition. Importantly, though we could, indeed, monitor actual effort/outcomes given the nature of the study’s design (as described in more detail below), workers were not aware of this fact.

The job itself consisted of three key sets of activities: a (required) main task with multiple parts, optional bonus tasks, and a worker survey administered at the end of the engagement to collect demographic information and exploratory insights into potential mechanisms/moderators. In both the main task and the bonus task, workers had both the opportunity and the incentive to engage in misconduct that, from their

---

9 While it would have been ideal to present the exact same amount of words/information in each of the treatment groups, it is in practice challenging to construct additional generic control group wording/information which does not bias our results in other ways. Though workers in the treatment groups were presented with more words/information to read than workers in the control group, we expect that any difference in time spent reading is unlikely to have biased our results upwards. First, workers in all of the treatment groups engaged in less misconduct than workers in the control group. If workers in the treatment groups were shirking more at the margin to gain time back (after being asked to read more), for example, this would bias our estimates of treatment effects towards zero – suggesting that our results are actually conservative rather than potentially overstated. Secondly, the amount of time required to the read the incremental treatment language is fairly small (a few seconds) vs. the amount of time it takes to actually complete any of the 5 individual required tasks (a couple of minutes on average).
perspective, would potentially go undetected by the employer. In particular, the extent to which the threat of monitoring could be perceived as credible varied between the main task and the optional bonus task, as the nature of the outputs requested varied between these two activities. We address this important distinction in more detail in the following section.

**Detailed Job Description**

Figure 3 presents a diagram illustrating the main task, which required workers to visit five interior designers’ websites and copy/paste a specified set of information into the “Contact” form on each site. (Figure 4 presents screenshots from one of these interior design websites.) Workers’ instructions were identical with the exception of language corresponding to the various treatment conditions (see Figure 3 for exact wording). Under typical conditions, it would be impossible for an employer to ascertain whether or not workers actually completed this task as instructed – thus our threat of monitoring treatment was arguably *not at all credible* from the worker’s perspective in this context. In actuality, though, we were the owners and operators of all five interior design websites and thus had access to all the data that workers were entering (or not entering). Moreover, upon hiring, in addition to being assigned a treatment condition, each worker was also assigned a unique random number. This number was inserted into the text that workers were to submit to each of the five websites (in the form of a telephone extension number; see Figure 3). This identifying number thus allowed us to ascertain which workers had completed the data entry task as instructed and which had shirked (i.e., not entered the required data into the “Contact” form for a given website).

*** Insert Figure 3 Here ***

*** Insert Figure 4 Here ***

After completing the main task for each of the five websites, workers were presented with an opportunity to complete an optional bonus task, detailed in Figure 5. These bonus tasks allowed for a different *form* of misconduct as well as for more opportunities to engage in misconduct in general (allowing for additional variation in misconduct at the intensive margin). This bonus task required workers to attempt
to contact the designer(s) by phone using contact information listed on the corresponding website(s). We provided workers with two scripts: one containing a market research survey to administer if a call recipient answered, and one to read if a call rang to voicemail (see Figure 5 for details; language corresponding to the SEV and EV treatments was again incorporated here). Bonus payment differed depending on the call outcome. We offered $0.25 for a completed market research survey (i.e., a maximum of $1.25 in bonus payment if responses were obtained from all five interior designers), and $0.10 for each voicemail left per our instructions. In practice, obtaining responses to the market research survey was impossible. As previously discussed, we were the owners of all of the interior designers’ websites. Accordingly, we also controlled all of the phone numbers listed on these websites, and during the course of this experiment, we ensured that no calls were answered. We could thus easily infer that any worker claiming to have obtained a response to the market research questions was misreporting this information.

We were also able to observe whether or not workers who claimed to have left a voicemail had actually done so. The same random number that allowed us to identify workers in the main task was also inserted into the voicemail script for the bonus task (once again, in the form of a telephone extension number; see Figure 5 for details). We were thus able to use this identifying number to discern whether or not workers had actually left voicemails in cases where claims were made.

*** Insert Figure 5 Here ***

What about the credibility of our monitoring threat in the context of the bonus task? Recall that workers in this condition were informed that 5% of HITs would be randomly selected to be reviewed for data quality. The voicemail outcome, like the central outcome in the main task, essentially involves workers leaving information on an external platform that a real employer would have no way of verifying. The higher-paying form of misconduct in the context of the bonus task, however, required the fabrication of market research survey responses. This is an outcome that (from the worker’s perspective) might plausibly be checked for accuracy – e.g., if multiple MTurk workers were hired to replicate the survey / validate responses for some subset of HITs. The threat of monitoring, then, was at least partially credible in this context.
Following completion of the main task (and bonus task(s), if applicable), workers were also asked to fill out a short survey. In this survey, we collected basic demographic information (gender, age, etc.) and asked exploratory questions aimed at assessing how workers’ agreement (on a five-point Likert scale) with statements about the employer and about misconduct in general might mediate or moderate any observed treatment effects.\textsuperscript{10} We presented this survey to workers as a way for us (as an employer) to learn more about our employees on MTurk, though it is possible that savvy workers might suspect at this point that they were actually participating in an academic study. However, because the survey was administered \textit{after} workers had already completed the actual portions of the job where misconduct was observed, we have no reason to believe that the nature of these questions influenced our key dependent variable in any way.

\textbf{Key Outcome Variables}

As discussed, a unique feature of our experimental setting was that we were able to cleanly observe various forms of employee misconduct, unbeknownst to the workers. In particular, we focus on two opportunities for misconduct in our empirical analysis: \textit{shirking in the main task} and \textit{fraudulent bonus claims}. In the bonus task, we are also able to distinguish between fraudulent voicemail claims (worth \$0.10 each) and fraudulent market research survey response claims (which were worth more at \$0.25 each, but which also required more effort to perpetrate since workers had to actually concoct answers to the market research questions in order to claim them).\textsuperscript{11}

The incentives that motivate workers to engage in these forms of misconduct arguably come down to money and time. Workers who make fraudulent bonus claims do so for a fairly obvious reason: direct financial gain. In contrast, the motivation to shirk is slightly more nuanced, though still likely related to financial gain, at least indirectly. Workers who shirk (i.e., those who report having completed a required

\textsuperscript{10} We do not find robust evidence of mediation and thus do not present any mediation results in this paper. We do, however, find suggestive evidence in support of our proposed theoretical arguments which we present as exploratory analyses at the end of our Results section.

\textsuperscript{11} To claim this bonus, workers had to actually fabricate responses to market research questions such as “Do you currently use any software products to help you manage project workflow and/or finances, and if so what products?” (Worker responses to the aforementioned question included answers such as “workflow software,” “Microsoft,” and “Acumatica.” The last of these is an actual small-business software solution.)
portion of the main task without having actually done so) ostensibly engage in this form of misconduct in order to finish the job (and receive their flat-rate payment) more quickly. In other words, workers who shirk do so to conserve *time* – which can then be spent earning money via other forms of work (on MTurk or elsewhere) or in leisure. The tradeoff for workers in both cases is the risk that the employer might detect their behavior and subsequently reject their submission – resulting in both a lost payment and reputational damage to their approval rating on MTurk.

**Key Explanatory Variables**

Our primary independent variables of focus are the randomized employer-level treatments previously discussed (EV, SEV, and threat of monitoring – as well as relevant interactions). In addition to these, we capture and include as controls in various specifications individual-level characteristics including worker gender, age, education, income, volunteer/donation activity, as well as data on workers’ (self-reported) relationship to the MTurk platform – specifically, the importance of the income that they earn on MTurk, their tenure on MTurk, and their MTurk rating.

**RESULTS**

Our raw sample consisted of 4,000 observations. After eliminating 93 observations where workers appear to have taken the survey multiple times (based on IP address and latitude/longitude data), our final sample consists of 3,907 observations. Note that for conciseness in presentation, we refer to the social/environmental values treatments as “SE values”, and to our threat of monitoring treatment as “monitoring” in what follows. Table 1 reports sample characteristics by condition and indicates that randomization across observable characteristics was successful, with no statistical differences between the treatment groups and the control group. Overall, workers in the sample are 54% female, with an average age of 35. Just over half the sample has a college degree or more advanced education. Roughly a quarter of workers have been engaged in gig work on MTurk for more than two years, and (though not reported in

---

12 In our pre-registration, we specified that we would continue to collect data either until we reached 4,000 observations or until 4 weeks had passed (whichever occurred first).

13 All results presented are substantively robust to the inclusion of these 93 omitted observations.
Table 1), nearly 90% report having been on the platform for at least one month. Most workers (76%) report their MTurk approval rating as 99% or higher. Table A1 in the appendix provides a correlation matrix for worker characteristics. All correlations are directionally sensible (e.g., workers with lower incomes indicate that the money they make on MTurk is more important to them, workers with college educations have higher incomes, etc.). This bolsters our confidence in the general integrity of the self-reported survey data and provides context for interpreting subsequent results.

*** Insert Table 1 Here ***

Figure 6 summarizes the overall incidence of worker misconduct (both shirking in the main task and fraudulent bonus claims) in our study. Roughly 80% of workers completed the main task in full, while 20% shirked at least one website entry. Among those who shirked, the distribution of outcomes is bimodal: most shirked either only once or in all five instances. Overall rates of fraudulent bonus claims are substantially lower, with fewer than 8% of workers engaging in this form of misconduct. However, bonus claims are markedly more common among workers who engage in shirking. Among workers who did not shirk any entries in the main task, only 2% claimed any fraudulent bonuses. This figure rises sharply to nearly 7% among workers who shirked once and to more than 52% among workers who shirked all five entry tasks.

*** Insert Figure 6 Here ***

Effects of Employer-Level Treatments

*** Insert Table 2 Here ***

Table 2 presents baseline differences in mean levels of misconduct by condition. In all instances, levels of misconduct are, as expected, lower in the treatment groups vs. the control group, though not all differences between groups are statistically significant. Tables 3 and 4 utilize regression analysis to examine the effects of our employer-level treatments more rigorously, and here statistical differences become more apparent – especially after worker controls are introduced to increase model precision. Our analysis focuses primarily on misconduct at the extensive margin (i.e., binary indicators for whether or not a worker engaged
in any misconduct), though results are similar when outcomes are instead specified as counts of the number of instances of shirking (from zero to five) or, in the case of bonuses, as a dollar value indicating the total amount of fraudulent bonuses claimed (see Models 4 and 8 in Tables 3 and 4). For ease of interpretation we utilize OLS in the specifications presented, but all results for binary outcomes are substantively robust to logit specifications.

Table 3 examines shirking in the main task (i.e., not entering specified information into the websites as instructed). Here, in the full sample (first four columns), all three employer-level treatments appear to have a negative effect on shirking, though only the organizational-values-oriented treatments (ethical values and SE values) are statistically significant in any specification. The last four columns of Table 3 replicate these results, but here we truncate our sample based on the time that workers spent viewing the page on which the employer-level treatments were initially presented. In particular, this sample omits workers in the bottom 10% of the distribution for time spent viewing the treatment language page – i.e., those who were less likely to have read the treatment language carefully or at all. Given that our treatments were administered solely via information presented in writing, we view this truncated sample of workers as a rough proxy for those workers who actually received our treatments. And, indeed, in this truncated sample, we find that the effects of all three individual treatments are larger and more statistically robust. Moreover, in both the full and truncated samples, the reduction in misconduct attributable to the treatments is meaningful in terms of economic magnitude. A coefficient of -0.03 in Model 2, for example, indicates that the proportion of workers who shirk falls by three percentage points (off a base of roughly 20%) when the

---

14 Recall that workers had five opportunities to claim bonuses, and in each instance could claim either a voicemail bonus (worth $0.10) or a survey bonus (worth $0.25). In analyses not presented, we also examine the effect of treatments on misconduct at the intensive margin – in other words, conditional on engaging in any misconduct, do treatments have any effect on the amount of misconduct a worker engages in? Here we find no statistically significant effects.

15 The “treatment language page” corresponds to the section depicted in Figure 2 where workers are initially presented with a condition-specific description of the hiring employer.
values-oriented treatments are utilized. This reduction in misconduct is more substantial (five to seven percentage points) in the truncated sample.

Table 4 mirrors Table 3, but focuses instead on misconduct in the bonus task. These results are, on the whole, more statistically robust than those for shirking in the main task. Here we see negative and statistically significant coefficients on the main effects for all three treatments in both the full sample and the truncated sample (though again, both effect size and statistical significance are greater in the truncated sample). The magnitude of these treatments is again economically significant. Roughly 8% of workers make fraudulent bonus claims in aggregate. Results in Table 4 indicate that the employer-level treatments we utilize each reduce the proportion of workers who make bonus claims by two to five percentage points (depending on the specification), lowering the base rate substantially.

Taken together, then, Tables 3 and 4 provide support for H1 (relating to communication of organizational values). With regard to H2 (relating to the threat of monitoring), results are slightly more complicated. Specifically, H2 posited that the threat of monitoring would reduce misconduct only in instances where this threat is credible. In actuality, results in Tables 3 and 4 suggest that the threat of monitoring may reduce misconduct even when this threat is entirely implausible. Indeed, results for the monitoring treatment are overall most robust (highly statistically significant across all specifications) in Table 4 – where the monitoring threat was at least partially credible. But coefficient estimates on the monitoring treatment are also statistically significant in at least some specifications even in Table 3, where our threat of monitoring was not at all credible. There are many potential explanations for why this might be. For one, it is possible that a large portion of workers do not fully think through whether or not the threat of monitoring is credible; rather, they simply take the threat at face value and modify their behavior accordingly. It could also be the case that the monitoring messaging has a subconscious effect (e.g., increasing the salience of rules, or of being caught and punished, or reducing workers’ feeling of autonomy) that results in behavioral changes even if workers conclude, rationally, that there is no way for the employer to act on this threat. It has been shown, for example, that feeling unconstrained by rules (Gino and Wiltermuth, 2014) and job autonomy is positively associated with individuals’ tendency to behave
unethically (Lu et al., 2017). Future research could further explore this somewhat counterintuitive finding in further depth.

Finally, results here provide some evidence in support of H3 (an interaction effect), particularly in the bonus task. In Table 4 in particular, positive and significant estimates on the interaction terms in Models 3, 4, 7, and 8 indicate that reductions in misconduct from organizational-values-oriented treatments and monitoring are not additive. For example, when monitoring is in effect, estimates in Models 7 and 8 suggest that there is little incremental benefit to implementing values-oriented policies, as coefficient estimates on main effects and interaction effects sum to roughly zero.

This interaction effect suggests that policies intended to better align workers’ interests with those of the firm by appealing to shared values may not be as successful when monitoring is also utilized. Figure 7 provides some suggestive evidence on the potential underlying mechanism for this interaction. On the left side of this figure, monitoring is not in effect. Here, we note that the ethical values and SE values treatments increase agreement with the statement, “I share this employer’s values” (vs. the control group) by a statistically significant margin. On the right side of the figure, when monitoring is in effect, the addition of values-oriented treatments do not change workers’ sense of shared values by a statistically significant amount. We posited previously that this might be because monitoring reduces trust between workers and the firm, causing it to be difficult for workers’ to form a sense of shared values. And, indeed, the bottom of Figure 7 indicates that workers in our experiment who were subject to monitoring were substantially less likely to agree with the statement, “This employer trusts its workers” than those who were not (72.3% vs. 78.1%, p<.001).16 This association is consistent with the theoretical underpinnings of H3, though we forthrightly caveat that our analysis is only exploratory in nature. We cannot conclusively rule out alternate explanations which could also contribute to the interaction, such as the possibility that any two treatments might be substitutes for one another. Future work could explore these mechanisms in more detail.

*** Insert Figure 7 Here ***

16 We emphasize that these findings are exploratory in nature and should be taken as suggestive rather than conclusive. A formal mediation analysis focused on a more precise understanding of mechanism is beyond the scope of this paper.
Individual Worker Characteristics and Misconduct

In addition to our employer-level treatments, Tables 3 and 4 both include a set of worker-level control variables in some models as indicated. While understanding the correlation between various worker characteristics and misconduct is not our primary focus, we do note that several characteristics are strongly correlated with misconduct and that observable individual characteristics explain a substantial portion of the variation in worker misconduct – between 20% and 37% depending on the specific outcome variable in question. We present key results in Table 5, with the caveat that these estimates should not be interpreted as causal (since individual characteristics are, of course, not randomly assigned).

*** Insert Table 5 Here ***

For the most part, correlational results here are unsurprising. High-performing workers (i.e., those with approval ratings of 99% or more) are substantially less likely to engage in misconduct than low-performing workers (those with approval ratings of less than 99%). Indeed, of the variables that we collected, workers’ MTurk rating is, perhaps unsurprisingly, the strongest individual-level predictor of misconduct. Women are also less likely than men to engage in misconduct, consistent with previous findings which have shown women to be less likely to cheat in the classroom, for example (Whitley et al., 1999). Interestingly, more highly educated individuals (those with at least a 4-year college degree) are most likely to engage in misconduct, as are those for whom the money earned on MTurk is more important.

Other statistically significant individual-level predictors may also be worth exploring more in future work, but we opt here to focus primarily on one particularly notable result: workers who report that they volunteer and donate more frequently\(^\text{17}\) are more likely to engage in misconduct (three percentage points more likely to shirk, and four percentage points more likely to fraudulently claim bonuses). This result initially seems counterintuitive, but it is actually in accordance with other findings – notably List and

\(^{17}\) The volunteer/donate score was constructed from responses to two individual questions in our survey: “Do you participate in volunteer work?” and “Do you donate to charity?” Workers could choose from four responses in each case, ranging from “Not at all,” assigned a value of 1 in our variable construction, to “Often” (for volunteering) or “Yes, a large amount” (for donations), assigned a value of 4 in our variable construction. The responses for these two questions were then summed to create an aggregate volunteer/donate score. The median of this composite variable was a score of 5 (on a scale ranging from 2 to 8). Our binary variable of focus assigns a value of 0 if a worker scores below 5 and a value of 1 if a worker scores a 5 or above.
Momeni (2020), who suggest that moral licensing may result in increased levels of worker misconduct. In other words, workers who feel that they are “doing good” in some ways may feel licensed to behave badly on the job. Specifically, List and Momeni posit that this mechanism may explain their finding that worker misconduct increases when firms engage in certain types of CSR.

Our finding that workers who volunteer and donate are more likely to engage in misconduct is in line with such a moral licensing mechanism. We thus explore this result more rigorously by conducting a follow-up survey with the same set of workers to determine its robustness. In particular, because responses to our initial survey were self-reported after workers had completed the main part of the task, it is reasonable to be concerned that perhaps causation might run in reverse – i.e., that engaging in misconduct caused workers to report higher levels of volunteer/donation activity, perhaps to compensate and make themselves feel better about having cheated.

We conducted our follow-up study in September 2019 (four months after our initial study) under the name of a different requester on MTurk so that workers would have no way of connecting this follow-up survey with our initial job. We invited all 3,907 workers from our initial study to complete the follow-up survey, for which we offered a payment of $0.50. We ultimately received 1,956 responses – just slightly more than a 50% response rate. Our follow-up survey asked workers to answer the same demographic questions that were included in our initial study, as well as some exploratory questions pertaining to their views on work, businesses/corporations, and lying. Our primary variables of interest, however, were those related to volunteer and donation history, since workers’ responses to these questions in the follow-up survey should not have been influenced in any direct way by their assigned condition or (mis)behavior in the initial job (i.e., previously discussed concerns of reverse causation are alleviated).

---

18 This follow-up study was pre-registered as an amendment to the initial pre-registration for the main study. The pre-registration is available from the authors upon request.

19 Note that the main tasks (and bonus tasks) assigned to workers in our original job were not part of our follow-up study (which contained only a survey). Because this follow-up survey required substantially less effort and time than our initial job, we payed workers less here than in our original study. In the follow-up survey, estimated completion time was three minutes, and workers’ actual median completion time was 1.7 minutes.

20 While we are able to effectively rule out reverse causation with this follow-up study, one possibility that we are not able to rule out is that of an omitted variable. Because data is self-reported, it is possible that workers who tend to engage in misconduct also tend to generally inflate their answers to questions about volunteer/donation history due to some other unobserved, underlying
Overall, responses were quite consistent across surveys. The correlation between workers’ composite volunteer/donate scores in the initial study (calculated from their originally reported answers) and their composite scores in the follow-up survey was 0.64. 84% of workers’ scores changed by no more than one point between surveys, and 96% of workers’ scores changed by no more than two points. Table A2 in the appendix compares coefficient estimates for regressions that utilize the original volunteer/donate variable to those that utilize an equivalent volunteer/donate variable calculated from follow-up survey data. Estimates are consistent across the original and follow-up data – bolstering confidence in our initially observed relationship between misconduct and volunteer/donation history. While not conclusive, these results strongly suggest that individual-level good behavior may, indeed, cause employees to feel licensed to behave badly. In contrast, our core finding that the communication of organizational values reduces misconduct suggests that employer-level “good behavior” does not elicit this same sort of moral licensing among workers. We discuss this nuance more in our concluding section.

Treatment Effect Heterogeneity Across Worker Types

On the whole, the Table 5 results discussed in the preceding section suggest that some types of workers are much more (or less) prone to misconduct than others. In this section we explore the extent to which our various treatments are effective on workers with a high propensity for misconduct vs. those with a low propensity for misconduct. We approach this analysis for both shirking and fraudulent bonuses as follows: 1) we fit logit models with (the relevant form of) misconduct as the dependent variable and all of the worker-level characteristics listed in Table 5 as predictors; 2) we use estimates from these models to generate predicted “propensity for misconduct” scores for all 3,907 workers; 3) we use these predicted misconduct scores to split our sample of workers into two groups: workers with a high propensity for

factor. Put differently, perhaps workers who are more likely to engage in misconduct are also more likely to lie in self-reported survey answers.
misconduct (those with propensity scores for misconduct in the top one-third of the distribution) and workers with medium/low propensity for misconduct (all remaining workers).21

It is not obvious what to expect ex ante with regard to where our treatments will have a stronger effect. On the one hand, “bad apple” workers with a high propensity for misconduct might be less receptive to values-oriented messaging and more skeptical of the threat of monitoring (or, alternatively less likely to read the task instructions in full – thus missing condition-specific language altogether). On the other hand, when the base rate of misconduct among workers is lower, there is simply less room for employer-level policies to further improve worker behavior (and vice versa). Table 6 presents results broken out across worker sub-samples. Here it is clear that for both shirking and fraudulent bonus claims, our employer-level treatments have a larger and more statistically significant effect on workers with a high propensity for misconduct than on workers with a medium/low propensity for misconduct. Indeed, these results suggest that there may be very little benefit to implementing employer-level policies like the ones we explore here when rates of worker misconduct are already low.

*** Insert Table 6 Here ***

Breakdown of Treatment Effects on Fraudulent Bonus Claims

Because the bonus task was optional, it is useful to understand how our various treatments may have affected two key outcomes separately: 1) selection into the bonus task itself, and 2) misconduct rates conditional on selection into the bonus task.22 Table 7 presents these more granular results for both the full sample and for the sub-sample of workers more likely to have read the treatment language (also utilized in parts of Tables 3 and 4). Results in both samples indicate that the threat of monitoring has a large and highly statistically significant effect on initial selection into the bonus condition (whereas the values-oriented treatments have none). Perhaps the most likely explanation for this result is that many workers who select into the bonus condition do so because they believe they can get away with cheating. The threat

---

21 Results are substantively robust to alternative splits in the distribution (e.g., top vs. bottom 50%, top 25% vs. bottom 75%, etc.), as well as to splits of the distribution into three distinct segments rather than two.

22 Note that conditional on selection into the bonus task, fraudulent claims occurred at a relatively high rate – roughly 50%.
of monitoring (at least partially credible in this context) changes this dynamic substantially, causing fewer workers who would have cheated to select into the bonus task in the first place. Other potential explanations may also be relevant (e.g., monitoring may demotivate workers, causing them to be less interested in completing “bonus” work), and future research could shed additional light on the specific mechanisms underlying selection patterns. Conditional on selection into the bonus task, Models 2-3 and 5-6 suggest that the effects of our employer-level treatments are most concentrated in reductions in fraudulent survey bonus claims (the higher-valued type of claim). In this outcome, we again find empirical support for H1, H2, and H3.

*** Insert Table 7 Here ***

CONCLUDING DISCUSSION

Employee misconduct is often challenging to observe empirically. Here, we present a novel real effort experiment revealing just how common misconduct can be in physically distant work settings where effort and outcomes are not (from the workers’ perspective) easily observable to the employer. In aggregate, more than 20% of workers in our study engaged in some form of misconduct on the job. The majority of this misconduct occurred in the form of shirking in the main task, where 6.9% of workers shirked data entry for one assigned website (out of five), and another 9.5% shirked all five data entry assignments. Many workers also fraudulently claimed bonuses for optional tasks that they had not actually completed. The cost of these fraudulent bonus claims was non-trivial, totaling 4% of our total wage bill paid to workers.

In traditional work contexts, there are many strategies that employers might utilize to curb such misconduct – the cultivation of a strong ethical company culture, for example (Pierce and Snyder, 2008). In the gig economy, however, many of these avenues tend to be unavailable due to the remote nature of work. We combine an agency perspective with a relationship-based behavioral perspective to develop theory around an alternative means by which gig and remote employers (and potentially traditional employers as well – though we recommend a healthy level of caution in the extrapolation of our results to traditional work contexts) can mitigate worker misbehavior: by communicating organizational values to the gig worker. In particular, we study communication of organizational values as a potential governance tool
for guarding against shirking and misreporting. Theoretically, by increasing the degree of interest alignment between the worker and employer, communication of organizational values should decrease the incidence of worker misconduct. And, indeed, in our empirical results, the inclusion of messaging that emphasized either ethical values or social/environmental values reduced the proportion of workers who engaged in misconduct by a statistically and economically significant margin. This finding suggests that, beyond pecuniary incentives which have been the focus of most extant research (e.g., Frank and Obloj 2014; Gubler et al. 2015; Larkin, 2014; Larkin and Pierce, 2015; Obloj and Sengul 2012; Oyer 1998) relationship-based motivators can serve as effective ways to improve interest alignment between principal and agent and thus mitigate adverse agent (employee) behavior (Flammer and Luo, 2017).

We also study the effect of a monitoring treatment. Though extant work has repeatedly shown direct monitoring to be effective in reducing misconduct (Becker, 1968; Hubbard, 2000; Nagin, Rebitzer, Sanders, and Taylor, 2002; Detert et al., 2007; Olken, 2007; DeHoratius and Raman, 2008; Pierce, Snow, and McAfee, 2015), our specific implementation differs slightly in that we focus instead on the mere threat of monitoring. Theoretically, we argue that this threat should only induce a behavioral reduction in misconduct when it is sufficiently credible to the worker. Our findings, however, suggest that the threat of monitoring can reduce misconduct even when the threat lacks credibility (though it is important to note that findings are more robust when the threat is credible vs. when it is not). Future work could explore whether even a non-credible threat of monitoring might have a subconscious effect on workers which increases the salience of rules, of being caught and punished, and/or reduces workers’ feeling of autonomy, for example. Indeed, it has been shown that feeling unconstrained by rules (Gino and Wiltermuth, 2014) and a sense of autonomy are positively associated with individuals’ tendency to behave unethically (Lu et al., 2017).

While the threat of monitoring alone reduces misconduct (ostensibly by decreasing the perception of information asymmetry between principal and agent), we also find that it diminishes the effectiveness of values-oriented messaging as a complementary tool. We provide suggestive evidence consistent with our theoretical argument that the threat of monitoring is likely to reduce trust between the employer and employee. We argue that this reduction in trust makes it more difficult for employees to develop the sense
of shared values (and corresponding improvements in interest alignment) that would otherwise occur as a result of these values-oriented policies. This finding is consistent with extant work which has shown that even in cases where monitoring is possible, it may have negative spillover effects (Aiello, 1993; Holland et al., 2015). More generally, it also points to the relevance of considering social and emotional influences (in addition to economic ones) when predicting the behavior of agents (Dunning et al., 2012). Indeed, our paper contributes to a stream of work seeking to combine agency theory with complementary perspectives to generate a more complex, realistic view of individual and organizational behavior (Eisenhardt, 1988; Anderson, 1985; Eccles, 1985; Flammer and Luo, 1998).

In addition, while not the primary focus of this paper, it is worth noting that our results indicate substantial variation in workers’ baseline propensity to engage in misconduct depending on individual-level characteristics. For researchers relying on MTurk samples, these differences shed light on the characteristics of participants who are most likely to cheat: those with lower MTurk approval ratings, who are more highly educated, who rely more heavily on MTurk as a source of income, and who are male (as opposed to female). Additionally, one finding in particular stands out: workers who report that they volunteer and donate more frequently actually engage in misconduct at higher rates.23 Taken together with our finding that communication of organizational values decreases worker misconduct, this result has potentially intriguing implications for the theoretical construct of moral licensing – particularly in the context of List and Momeni (2020), who find a (weakly) positive relationship between CSR and rates of shirking among workers.

The theoretical argument put forth by List and Momeni is that CSR should elicit moral licensing and thus increase worker misconduct. Notably, though, the authors’ main results are statistically significant only when the pro-social act is framed in direct relation to the individual employee.24 Consistent with this notion, our results around (self-reported) worker volunteer/donation habits and misconduct (bolstered by

---

23 It is possible that individuals who are more prone to engaging in misconduct might also be more likely to lie in general (and thus, misrepresent or have an exaggerated view of their charitable behavior).

24 Specifically, only when workers are told that, “We donate the equivalent of x% of our wage bill in cash (on behalf of all workers who help us with this project) to UNICEF Education Programs” do List and Momeni (2017) find a statistically significant effect on unethical behavior. When the same statement without the parenthetical is administered, key results are insignificant. One important open question, then, is whether or not moral licensing is elicited when CSR is framed in terms of the organization rather than the individual.
evidence from a follow-up survey) suggest that moral licensing may indeed be relevant when employees who behave well on an individual basis in one context (volunteering/donating) feel licensed to behave badly in another (misconduct in the workplace). A key distinction here is that the act that elicits moral licensing occurs at the same level (empirically, at the same unit of observation) as the adverse behavior of focus. Indeed, this is also in line with extant work which suggests that good behavior at the organization level may result in adverse behavior at the organization level, with Luo, Kaul, and Seo (2018) showing evidence that firms that donate more have more subsequent oil spills, for example.

Our findings suggest that individual-level good behavior does seem to appear to make workers feel licensed to behave badly as individuals. In contrast, if it is the organization that is framed as “doing good” (as is the case in our values-oriented treatments), individual-level adverse behavior does not appear to ensue. Put simply, our results suggest that as long as social responsibility is framed at the firm-level (as opposed to the individual level), it should decrease (rather than increase) misconduct. This distinction is nuanced but also appears to be quite important. Indeed, our results here, taken in concert with those of List and Momeni (2020), suggest that there may, more generally, be substantial variation with respect to the way in which different types of CSR affect worker misconduct (and likely worker behavior more broadly). The specific way in which CSR is implemented and communicated in practice can, seemingly, have a large effect on the magnitude and even the direction of its impact. A more thorough understanding of this nuance is critical for scholars, who are increasingly recognizing that the broader category of “CSR” needs to be deconstructed into its various parts to understand the mechanisms through which various socially responsible policies and activities can benefit the firm (Burbano, Mamer, and Snyder, 2018).

In conclusion, the theory that we develop around communication of values and misconduct is highly relevant for employers and platforms in the growing gig economy, as well as in increasingly prevalent remote work contexts. While we forthrightly acknowledge that MTurk has important limitations as a setting for studying issues pertaining to traditional work contexts, we posit that this setting is actually ideal for studying misconduct in the gig economy, as MTurk jobs fit all of the criteria for prototypical gig work: jobs are completed while physically distant from the employer, and are furthermore task-based, (Kaine and
Josserand, 2019) short-term, and facilitated by an intermediary platform that connects the requesters with the workers. Furthermore, as MTurk workers completing short-term jobs are arguably less likely to develop a sense of obligation or connection to their employer (compared with workers in relatively longer-term gig or remote work contexts), we would expect this to bias our results downwards. The effects of communicating organizational values and the interaction effect of such communication with the threat of monitoring are thus likely to be even greater in gig or remote contexts which are longer-term. Future work might examine how task/engagement duration and perceived distance from the employer (Burbano 2019) shape the efficacy of various policies aimed at reducing misconduct.

While our empirical results are drawn from MTurk specifically, the mechanisms underlying the policies that we study should theoretically extend to other types of gig work (e.g., point-to-point transport and food delivery services) and to remote work. These mechanisms should also likely apply in many traditional work contexts, though here it is not clear how important these mechanisms might be in relation to other factors such as organizational culture, and we recommend a healthy level of caution in the extrapolation of our results to traditional work settings. Future work which examines employee misconduct in other contexts will serve as useful complements to this research.

Finally, given the challenge of observing misconduct in practice (as well as that of establishing a causal relationship between organization-level characteristics and individual-level misconduct), our innovative experimental design is also an important contribution – allowing us to accurately observe and measure employee misconduct in a natural work context and to examine the causal effects of various organization-level policies on this outcome. There is substantial value in a study like ours, which focuses on the estimation of treatment effects and is intentionally designed to eliminate the possibility that omitted variables or differences in worker selection patterns confound results. That said, it is also important to consider the fact that individual-level characteristics appear to explain a substantial portion of the observed variation in misconduct across workers in our data – indeed, substantially more of the variation than our treatment effects explain. This implies that attracting and selecting the right workers is also an important component of any broader organizational strategy for mitigating misconduct. Future work might thus
examine the way in which various organization-level policies affect the selection of different types of workers into different opportunities and, subsequently, how these differences in selection patterns influence corresponding misconduct outcomes.

References
Burbano VC and Ostler J (2020) Differences in consumer-benefiting misconduct by nonprofit, for-profit, and public organizations. *Organizational Behavior and Human Decision Processes*.
Cialdini RB (1996) Social influence and the triple tumor structure of organizational dis-honesty. In
Messick & AE Tenbrunsel (Eds.), Codes of conduct: Behavioral research into business ethics (pp. 44–58). New York: Russell Sage.


Gino F, Pierce L. (2010b) Lying to level the playing field: Why people may dishonestly help or hurt others to create equity,” *Journal of Business Ethics* 95(1): 89–103.


Wiltermuth SS (2011) Cheating more when the spoils are split. *Organizational Behavior and Human

### Table 1: Sample Characteristics by Condition (Randomization Balance)

<table>
<thead>
<tr>
<th></th>
<th>Only Ethical</th>
<th>Only Values †</th>
<th>Only SE Values ††</th>
<th>Ethical Values + Monitoring ††</th>
<th>SE Values †† + Monitoring ††</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>3,907</td>
<td>635</td>
<td>632</td>
<td>692</td>
<td>667</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>0.54</td>
<td>0.53</td>
<td>0.54</td>
<td>0.52</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>35.3</td>
<td>35.2</td>
<td>35.4</td>
<td>35.7</td>
<td>35.3</td>
</tr>
<tr>
<td><strong>4-Yr College Degree</strong></td>
<td>0.54</td>
<td>0.56</td>
<td>0.52</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td><strong>Income ≥ $50K</strong></td>
<td>0.51</td>
<td>0.53</td>
<td>0.49</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td><strong>Volunteer/Donate ≥ Sample Median</strong></td>
<td>0.54</td>
<td>0.53</td>
<td>0.51</td>
<td>0.53</td>
<td>0.55</td>
</tr>
<tr>
<td><strong>MTurk Income Important (Likert)</strong></td>
<td>3.57</td>
<td>3.57</td>
<td>3.51</td>
<td>3.60</td>
<td>3.58</td>
</tr>
<tr>
<td><strong>Active on MTurk &gt; 2 Years</strong></td>
<td>0.28</td>
<td>0.27</td>
<td>0.30</td>
<td>0.27</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>MTurk Approval Rate ≥ 99%</strong></td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
<td>0.76</td>
<td>0.77</td>
</tr>
</tbody>
</table>

**Note:** All variables are binary except Age (which takes integer values) and MTurk Income Important, which is measured on a 5-point Likert Scale (5 = strongly agree that money earned on MTurk is an important source of income). Standard deviations reported in parentheses. Asterisks indicate sample means that are statistically different from the control group. (*** p < 0.01, ** p < 0.05, * p < 0.1)

† Social/environmental values.

†† For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement.

---

### Table 2: Baseline Levels of Misconduct by Condition

<table>
<thead>
<tr>
<th></th>
<th>Only Ethical</th>
<th>Only Values †</th>
<th>Only SE Values ††</th>
<th>Ethical Values + Monitoring ††</th>
<th>SE Values †† + Monitoring ††</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shirked Any Websites?</strong></td>
<td>0.195</td>
<td>0.204</td>
<td>0.188</td>
<td>0.198</td>
<td>0.177†</td>
</tr>
<tr>
<td><strong>Fraudulently Claimed Any Bonuses?</strong></td>
<td>0.079</td>
<td>0.087</td>
<td>0.069**</td>
<td>0.069**</td>
<td>0.069***</td>
</tr>
<tr>
<td><strong>Avg. # of Websites Shirked</strong></td>
<td>0.635</td>
<td>0.647</td>
<td>0.622</td>
<td>0.553</td>
<td>0.619</td>
</tr>
<tr>
<td><strong>Avg. # of Fraudulent Bonus Claims</strong></td>
<td>0.248</td>
<td>0.257</td>
<td>0.213**</td>
<td>0.213**</td>
<td>0.213*</td>
</tr>
<tr>
<td><strong>Avg. Total $ Value of Fraudulent Bonus Claims</strong></td>
<td>0.039</td>
<td>0.040</td>
<td>0.032**</td>
<td>0.033**</td>
<td>0.035*</td>
</tr>
</tbody>
</table>

**Note:** The first two variables are binary indicators for whether or not a worker engaged in any misconduct as specified. Workers could shirk between 0 and 5 websites and fraudulently claim between 0 and 5 bonuses (worth up to $1.25 in fraudulent claims). Asterisks indicate sample means that are statistically different from the control group. (*** p < 0.01, ** p < 0.05, * p < 0.1)

† Social/environmental values.

†† For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement.
## Table 3: Treatment Effects on Shirking

|                     | Full Sample |                     |                      | Workers More Likely to Have Read Treatment Language  
|---------------------|-------------|---------------------|---------------------|--------------------------------------------------
|                     | Any Websites Entries Shirked? (Binary) | # Websites Shirked (Count 0-5) | Any Websites Entries Shirked? (Binary) | # Websites Shirked (Count 0-5) |
|                     | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Ethical Values      | -0.02 | -0.03* | -0.03 | -0.05 | -0.03** | -0.03** | -0.05** | -0.16* |
|                     | (0.02) | (0.01) | (0.02) | (0.07) | (0.02) | (0.02) | (0.03) | (0.09) |
| SE Values†          | -0.02 | -0.03** | -0.05** | -0.10 | -0.04** | -0.04*** | -0.07*** | -0.21** |
|                     | (0.02) | (0.01) | (0.02) | (0.07) | (0.02) | (0.02) | (0.02) | (0.08) |
| Monitoring††         | -0.02 | -0.02 | -0.02 | -0.03 | -0.03** | -0.03** | -0.06** | -0.16* |
|                     | (0.01) | (0.01) | (0.02) | (0.07) | (0.01) | (0.01) | (0.02) | (0.09) |
| Monitoring†† x Ethical Values | -0.00 | -0.10 | 0.03 | (0.03) | (0.10) | (0.03) | (0.11) |
| Monitoring†† x SE Values† | 0.02 | 0.00 | 0.05* | (0.03) | (0.09) | (0.03) | (0.11) |
| Controls            | No | Yes | Yes | Yes | No | Yes | Yes | Yes |
| Constant            | 0.22*** | 0.66*** | 0.66*** | 4.04*** | 0.24*** | 0.69*** | 0.71*** | 4.40*** |
|                     | (0.01) | (0.10) | (0.10) | (0.41) | (0.02) | (0.11) | (0.11) | (0.44) |
| Observations        | 3,907 | 3,907 | 3,907 | 3,907 | 3,517 | 3,517 | 3,517 | 3,517 |
| R-squared           | 0.00 | 0.20 | 0.20 | 0.37 | 0.00 | 0.21 | 0.21 | 0.38 |

Robust standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1

**Note:** All regressions are OLS (though all results for binary outcomes are substantively robust to logit specifications). Models 2-4 and 6-8 include the worker-level control variables from Table 1 (Female, Age ≥ 4-Yr College, Income ≥ $50K, Vol/Don ≥ Median, MTurk $ Important, On MTurk ≥ 2 Yrs, MTurk Approval ≥ 99%) as well as the following variables: the natural log of the total seconds the worker took to complete the job, the natural log of the total seconds the worker spent viewing the treatment language page, an indicator for whether or not the worker passed an attention check question, indicators for political affiliation and military service, and indicators for inconsistencies in the data (ZIP code and geo-coordinate mismatches and gender mismatches on primary and follow-up surveys).

† Social/environmental values.

†† For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement. Notably, here in the main task, this threat was not actually credible from the perspective of the worker.

---

*a* This sample omits workers in the bottom 10% of the distribution for time spent viewing the page containing condition-specific language.

---

Note: All regressions are OLS (though all results for binary outcomes are substantively robust to logit specifications). Models 2-4 and 6-8 include the worker-level control variables from Table 1 (Female, Age ≥ 4-Yr College, Income ≥ $50K, Vol/Don ≥ Median, MTurk $ Important, On MTurk ≥ 2 Yrs, MTurk Approval ≥ 99%) as well as the following variables: the natural log of the total seconds the worker took to complete the job, the natural log of the total seconds the worker spent viewing the treatment language page, an indicator for whether or not the worker passed an attention check question, indicators for political affiliation and military service, and indicators for inconsistencies in the data (ZIP code and geo-coordinate mismatches and gender mismatches on primary and follow-up surveys).

† Social/environmental values.

†† For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement. Notably, here in the main task, this threat was not actually credible from the perspective of the worker.
Table 4: Treatment Effects on Fraudulent Bonus Claims

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th></th>
<th></th>
<th>Workers More Likely to Have Read Treatment Language a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Any Fraudulent Bonuses</td>
<td>Value of Fraudulent Claims</td>
<td>Any Fraudulent Bonuses</td>
<td>Value of Fraudulent Claims</td>
</tr>
<tr>
<td></td>
<td>(Binary)</td>
<td>(Dollars)</td>
<td>(Binary)</td>
<td>(Dollars)</td>
</tr>
<tr>
<td>Ethical Values</td>
<td>−0.01</td>
<td>−0.02</td>
<td>−0.03**</td>
<td>−0.02**</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>SE Values†</td>
<td>−0.01</td>
<td>−0.02*</td>
<td>−0.03*</td>
<td>−0.02**</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Monitoring††</td>
<td>−0.02***</td>
<td>−0.02***</td>
<td>−0.04***</td>
<td>−0.03***</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Monitoring†† x Ethical Values</td>
<td>0.04*</td>
<td>0.02</td>
<td>0.05**</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Monitoring†† x SE Values†</td>
<td>0.02</td>
<td>0.02*</td>
<td>0.03</td>
<td>0.04**</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Constant</td>
<td>0.10***</td>
<td>0.23***</td>
<td>0.24***</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Observations</td>
<td>3,907</td>
<td>3,907</td>
<td>3,907</td>
<td>3,907</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.00</td>
<td>0.23</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

a This sample omits workers in the bottom 10% of the distribution for time spent viewing the page containing condition-specific language.

Note: All regressions are OLS (though all results for binary outcomes are substantively robust to logit specifications). Models 2-4 and 6-8 include the worker-level control variables from Table 1 (Female, Age ≥ 4 Yr College, Income ≥ $50K, Vol/Don ≥ Median, MTurk $ Important, On MTurk > 2 Yrs, MTurk Approval ≥ 99%) as well as the following variables: the natural log of the total seconds the worker took to complete the job, the natural log of the total seconds the worker spent viewing the treatment language page, an indicator for whether or not the worker passed an attention check question, indicators for political affiliation and military service, and indicators for inconsistencies in the data (ZIP code and geo-coordinate mismatches and gender mismatches on primary and follow-up surveys).

† Social/environmental values.

†† For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement. Here, in the bonus task, this threat was (at least partially) credible from the perspective of the worker.
Table 5: Individual Worker Characteristics and Misconduct

<table>
<thead>
<tr>
<th></th>
<th>Shirking Binary</th>
<th>Shirking Count (0-5)</th>
<th>Fraudulent Bonus Claims Binary</th>
<th>Fraudulent Bonus Claims Value ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Female</td>
<td>−0.04***</td>
<td>−0.18***</td>
<td>−0.01*</td>
<td>−0.01**</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Age†</td>
<td>−0.00</td>
<td>0.00</td>
<td>−0.01***</td>
<td>−0.01***</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>≥ 4-Yr College</td>
<td>0.04***</td>
<td>0.20***</td>
<td>0.04***</td>
<td>0.02***</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Income ≥ $50K</td>
<td>−0.00</td>
<td>−0.03</td>
<td>−0.02**</td>
<td>−0.01</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Vol/Don ≥ Median</td>
<td>0.03***</td>
<td>0.16***</td>
<td>0.04***</td>
<td>0.02***</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>MTurk $ Important</td>
<td>0.02***</td>
<td>0.08***</td>
<td>0.02***</td>
<td>0.01***</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>On MTurk &gt; 2 Yrs</td>
<td>0.03 *</td>
<td>−0.03</td>
<td>−0.00</td>
<td>−0.00</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>MTurk Approval ≥ 99%</td>
<td>−0.12***</td>
<td>−0.55***</td>
<td>−0.05***</td>
<td>−0.04***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.06)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.63***</td>
<td>3.94***</td>
<td>0.21***</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.41)</td>
<td>(0.08)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Observations</td>
<td>3,907</td>
<td>3,907</td>
<td>3,907</td>
<td>3,907</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.20</td>
<td>0.37</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1

Note: All regressions are OLS (though all results for binary outcomes are substantively robust to logit specifications). All models also include the following worker-level control variables: the natural log of the total seconds the worker took to complete the job, the natural log of the total seconds the worker spent viewing the treatment language page, an indicator for whether or not the worker passed an attention check question, indicators for political affiliation and military service, and indicators for inconsistencies in the data (ZIP code and geo-coordinate mismatches and gender mismatches on primary and follow-up surveys).

† The age variable is scaled down by a factor of ten to make coefficients more interpretable.
Table 6: Treatment Effects on Rates of Shirking and Fraudulent Bonus Claims

Comparison of Workers with High vs. Medium/Low Propensity to Engage in Misconduct

<table>
<thead>
<tr>
<th></th>
<th>Any Websites Entries Shirked?</th>
<th>Any Fraudulent Bonuses Claimed?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Workers with High Propensity for Misconduct (1)</td>
<td>Workers with Medium/Low Propensity for Misconduct (2)</td>
</tr>
<tr>
<td>Ethical Values</td>
<td>−0.03 (0.02)</td>
<td>0.01 (0.02)</td>
</tr>
<tr>
<td></td>
<td>−0.07** (0.04)</td>
<td>0.02 (0.02)</td>
</tr>
<tr>
<td>SE Values †</td>
<td>−0.05** (0.02)</td>
<td>−0.09** (0.04)</td>
</tr>
<tr>
<td></td>
<td>−0.02 (0.02)</td>
<td>−0.02 (0.02)</td>
</tr>
<tr>
<td>Monitoring ††</td>
<td>−0.00 (0.03)</td>
<td>−0.00 (0.06)</td>
</tr>
<tr>
<td></td>
<td>−0.00 (0.04)</td>
<td>−0.00 (0.06)</td>
</tr>
<tr>
<td>Monitoring †† x Ethical Values</td>
<td>0.02 (0.03)</td>
<td>0.06 (0.06)</td>
</tr>
<tr>
<td></td>
<td>0.01 (0.03)</td>
<td>0.01 (0.06)</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Constant</td>
<td>0.66*** (0.10)</td>
<td>1.27*** (0.16)</td>
</tr>
<tr>
<td></td>
<td>0.24 (0.15)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>3,907</td>
<td>2,606</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.20</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: All regressions are OLS (though results are substantively robust to logit specifications). All models include the same control variables listed in Tables 3-5.

† Social/environmental values.
†† For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement.
Table 7: Breakdown of Treatment Effects on Fraudulent Bonus Claims

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th></th>
<th>Workers More Likely to Have Read Treatment Language</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Worker Selected Into Any Bonus Opportunities?</td>
<td>Voicemails ($0.10)</td>
<td>Survey Answers ($0.25)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Ethical Values</td>
<td></td>
<td>−0.00</td>
<td>−0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>SE Values†</td>
<td>−0.01</td>
<td>(0.02)</td>
<td>−0.09</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.05)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Monitoring††</td>
<td>−0.05***</td>
<td>(0.02)</td>
<td>−0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.07)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Monitoring†† x Ethical Values</td>
<td>0.03</td>
<td>0.05</td>
<td>0.14*</td>
</tr>
<tr>
<td>Monitoring†† x SE Values†</td>
<td>0.03</td>
<td>−0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Constant</td>
<td>−0.14</td>
<td>1.17***</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.26)</td>
<td>(0.19)</td>
</tr>
<tr>
<td>Observations</td>
<td>3,907</td>
<td>576</td>
<td>576</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.21</td>
<td>0.09</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p < 0.01, ** p < 0.05, * p < 0.1

a. This sample omits workers in the bottom 10% of the distribution for time spent viewing the page containing condition-specific language.

Note: All regressions are OLS (though results are substantively robust to logit specifications). All models include the same control variables listed in Tables 3-5.

† Social/environmental values.

†† For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement.
Figure 1: Three-by-Two Experimental Design

<table>
<thead>
<tr>
<th>No Monitoring</th>
<th>Ethical Values</th>
<th>SE Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Monitoring</td>
<td>C1: Control</td>
<td>C3: Only SE Values</td>
</tr>
<tr>
<td>Monitoring</td>
<td>C2: Only Ethical Values</td>
<td>C5: Ethical Values + Monitoring</td>
</tr>
<tr>
<td></td>
<td>C4: Only Monitoring</td>
<td>C6: SE Values + Monitoring</td>
</tr>
</tbody>
</table>

Note: “SE Values” = Social/environmental values. For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement.
Note: “SE Values” = Social/environmental values. For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement.
Figure 3: Main Task Flow

|-------------|---------------|--------------------|-------------------------------|--------------|--------------------------|
| Website X of 5  
www.[designer_website_name].com  
Please copy and paste the above link into a separate browser window and enter the following information into the "Contact" section of the website (if available).  
| Please remember that a portion of HITs will be randomly selected for data quality review by another worker.  |

If contact form asks for an email address: [XX@XXXXXXXXX.com]  
If contact form asks for a name: [OUR PROVIDED NAME HERE]  
If contact form asks for a subject: Interior Design Software  
If contact form includes an open entry space, copy and paste in this message:  

Do you need a tool to help you organize the finances of your design jobs? We are a startup offering a software product tailored specifically to meet the needs of interior designers.  
Our company believes in maintaining a culture of accountability and transparency. We strive to be honest, ethical, and fair in all that we do.  
We would love the opportunity to learn more about your potential software needs. Please contact us by phone at XXX-XXX-XXXX, extension [BANG_ID] for more information and special promotional pricing offers!  
We would love the opportunity to learn more about your potential software needs. Please contact us by phone at XXX-XXX-XXXX, extension [BANG_ID] for more information and special promotional pricing offers!  
We would love the opportunity to learn more about your potential software needs. Please contact us by phone at XXX-XXX-XXXX, extension [BANG_ID] for more information and special promotional pricing offers!  
Was the website functioning? [y/n]  
Was there a "Contact" page on the website? [y/n]  
If there was a "Contact" page, did you enter the above information and submit the information on the "Contact" page? [y/n]  
If you selected "No" in the question above, please elaborate: [optional text response]  

After completing the basic task for each website, workers are asked if they wish to complete an optional bonus task (see Figure 5 for details)  

Note: “SE Values” = Social/environmental values. For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement.
Figure 4: Screenshots from One of Our Interior Design Websites

(a) Home Page

(b) Contact Page
Figure 5: Bonus Task Flow

The optional bonus task requires workers to contact the designer(s) by phone

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Workers receive an overview of bonus task</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workers decide if they wish to complete the bonus task</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workers who opt in receive a script for their calls to the design firm(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workers are required to report the task outcome (includes a chance to opt out)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If there is a phone number on the website, we will pay you a bonus if you call it and deliver a short 4-question market research survey (or leave a voicemail as instructed).

Website X of 5

www.[designer_website_name].com

You will receive a $0.25 bonus if you are able to obtain answers to our market research survey. You will receive a $0.10 bonus if you are not able to speak to anyone but leave a voicemail as instructed.

Do you wish to complete this optional bonus assignment?

| Worker selects YES |
| Worker selects NO |

Instructions for Bonus Call:

If someone answers, use this script:

"Hi, I'm calling with 4 short questions about your design business software needs. Would you be willing to answer these questions? This will take 1 minute of your time."

How many residential clients do you serve per year, on average?

How many employees do you have, including yourself?

Do you currently use any software products to help you manage project workflow and/or finances? [If so, what products?]

Would you be interested in a new software product to help you manage project workflow and/or finances?

If no one answers, leave this voicemail:

"Hello, I'm calling on behalf of XXXXX Software. We have developed a tool for designers and would love the opportunity to learn more about your potential software needs.

Our company believes in maintaining a culture of accountability and transparency. We strive to be honest, ethical, and fair in all that we do.

Please contact us by phone at XXX-XXXX-XXXX, extension [RND_ID] for more information and special promotional pricing offers!"

"Hello, I'm calling on behalf of XXXXX Software. We have developed a tool for designers and would love the opportunity to learn more about your potential software needs.

Our company believes in giving back to the community and in improving the environment. We donate 5% of our profits to charitable organizations each year and invest in environmentally sustainable practices.

Please contact us by phone at XXX-XXXX-XXXX, extension [RND_ID] for more information and special promotional pricing offers!"

What was the call outcome?

Changed my mind and decided not to complete this bonus task (no bonus)

Obtained answers to the market research questions ($0.25 bonus; must fill out answers below)

Left a voicemail as instructed ($0.10 bonus)

Other (please explain to determine eligibility for bonus)

[Workers who selected the second option above are required to input answers to the market research questions here]

Workers are taken to the main task instructions page for the next website (or to the final worker survey after completing the fifth website)

Note: "SE Values" = Social/environmental values. For brevity, we use the term "monitoring," but in all cases it is only the threat of monitoring that we actually implement.
Figure 6: Overall Incidence of Worker Misconduct

<table>
<thead>
<tr>
<th>Website Entries Shirked</th>
<th>Percent of Workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>80.5%</td>
</tr>
<tr>
<td>1</td>
<td>6.9%</td>
</tr>
<tr>
<td>2</td>
<td>1.3%</td>
</tr>
<tr>
<td>3</td>
<td>1.0%</td>
</tr>
<tr>
<td>4</td>
<td>0.9%</td>
</tr>
<tr>
<td>5</td>
<td>9.5%</td>
</tr>
</tbody>
</table>

| % of Workers Fraudulently Claiming Bonuses | 2.3% | 6.7% | 8.2% | 18.4% | 34.3% | 52.2% | **7.9%** |

*Full Sample*
For both shared values and trust, workers were asked about the extent to which they agreed with the corresponding statements on a 5-point Likert scale (1=strongly disagree, 5= strongly agree). In the data above, workers who responded with a 4 or 5 on this scale are included as “agreeing.” Asterisks indicate sample means that are statistically different from the relevant control group. (** p<0.01, * p<0.05, * p<0.1)

Note: “SE Values” = Social/environmental values. For brevity, we use the term “monitoring,” but in all cases it is only the threat of monitoring that we actually implement.
1 Appendix: Additional Tables and Figures

Table A1: Correlation Matrix for Individual-Level Worker Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Age</th>
<th>≥ 4-Yr College</th>
<th>Income ≥ $50K</th>
<th>Vol/Don ≥ Median</th>
<th>MTurk $ Important</th>
<th>On MTurk &gt; 2 Yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td>0.09***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 4-Yr College</td>
<td></td>
<td>-0.09***</td>
<td>-0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income ≥ $50K</td>
<td></td>
<td>-0.01</td>
<td>0.05***</td>
<td>0.28***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vol/Don ≥ Median</td>
<td>0.11***</td>
<td>0.11***</td>
<td>0.14***</td>
<td>0.14***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTurk $ Important</td>
<td>0.09***</td>
<td>-0.02</td>
<td>-0.17***</td>
<td>-0.25***</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>On MTurk &gt; 2 Yrs</td>
<td>0.00</td>
<td>0.17***</td>
<td>-0.00</td>
<td>0.01</td>
<td>-0.06***</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>MTurk Approval ≥ 99%</td>
<td>-0.04***</td>
<td>0.08***</td>
<td>0.01</td>
<td>0.08***</td>
<td>-0.07***</td>
<td>-0.02</td>
<td>0.14***</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1

Table A2: OLS Coefficient Estimates on the Binary Volunteer/Donate Variable

Original vs. Follow-Up Survey Data

<table>
<thead>
<tr>
<th></th>
<th>Any Website Entries Shirked?</th>
<th>Any Bonuses Claimed?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original Data (1)</td>
<td>Follow-Up Data (2)</td>
</tr>
<tr>
<td>Vol/Don ≥ Median</td>
<td>0.028*</td>
<td>(0.016)</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Observations</td>
<td>1.956</td>
<td>1.956</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.20</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: All models also contain all of the variables included in Table 5. Models 1 and 3 here mirror the original specifications 1 and 3 in Table 5 exactly, except for the fact that results here are estimated using only the 1,956 workers who also completed the follow-up survey rather than the full sample of 3,907 workers. Models 2 and 4 replace the original volunteer/donate data with corresponding data collected in the follow-up survey. There is very little difference in results obtained using the original data vs. those obtained using the follow-up data (if anything, the initially observed positive correlation between volunteer/donation activity and misconduct is actually stronger when follow-up data is utilized).