The Effects of Investigative Sanctioning Systems on Wrongdoing, Reporting, and Helping: A Multi-Party Perspective

Celia Chui[a*] & Manuel Grieder[b,c]

*: Corresponding author

a: Questrom School of Business, Boston University, 595 Commonwealth Avenue, Boston, MA 02215
Email: cchui@bu.edu
Phone: 1-857-234-1269

b: ETH Zurich, Dep. Of Humanities, Social, and Political Science, Clausiusstrasse 37, 8092 Zurich, Switzerland
Email: manuel.grieder@econ.gess.ethz.ch

c: Zurich University of Applied Sciences (ZHAW), School of Management and Law, Bahnhofplatz 12, 8400 Winterthur, Switzerland
Email: manuel.grieder@zhaw.ch

(*Forthcoming* at Organization Science)

Acknowledgements

The authors are very grateful to Senior Editor Steven Blader as well as three anonymous reviewers for providing us with their valuable guidance and constructive feedback throughout the reviewing process. For helpful advice and support on this project, we thank Joerg Dietz, Maryam Kouchaki, Brian Lucas, Martha Maznevski, the late Keith Murnighan, Renate Schubert, Daniel Waeger, and Christian Zehnder; participants in seminars at the University of Lausanne and Northwestern University; and participants at the 2014 Kellogg-Booth Student Symposium, the 2015 Academy of Management Meetings, and the 2015 International Association for Conflict Management Conference. For research assistance, we thank Simon Frischknecht and Franziska Isle.

ABSTRACT

Over the past two decades, organizations have established sanctioning systems as an important component of their ethical infrastructures to detect and punish wrongdoing. However, empirical knowledge about the overall effectiveness of such systems remains limited. Existing studies have mostly adopted a single-party perspective, even though many wrongdoing situations involve dynamic multi-party interactions between actors, recipients and observers of wrongdoing. Moreover, most existing research has emphasized an economic perspective – that sanctioning systems only affect behavior because of economic considerations while crowding out ethical ones. In this research, we develop a moral and normative perspective of sanctioning systems. Using a novel experimental game design, our study focuses on the investigative dimension of sanctioning systems to examine their psychological and behavioral effects in actor-recipient-observer wrongdoing *interactions*. Findings reveal that investigative sanctioning systems influence wrongdoing, reporting, and helping decisions as well as alter ethical and normative considerations, such that as systems become stronger, wrongdoing behaviors are judged as more unethical and perceived as less typical than when weaker systems are in place. These moral judgments and norm perceptions mediate the effect of investigative sanctioning system strength on wrongdoing behavior. Our research extends previous empirical and theoretical work on sanctioning systems by applying a more holistic perspective and by demonstrating that highly effective systems can serve as important behavioral guides because they activate and alter moral and normative considerations about wrongdoing.

Keywords: Sanctioning systems, behavioral ethics, moral judgment, social norms, wrongdoing, reporting, third-party helping, false accusations

In recent years, companies have increasingly established formal systems and investigatory resources to help detect and discipline employee misconduct (Smith-Crowe et al. 2015; Tenbrunsel et al., 2003). According to the 2018 Global Business Ethics Survey, 54% of employees worldwide reported that their organizations had internal systems to investigate and discipline ethical violations (Ethics and Compliance Initiative (ECI) 2018), suggesting their prevalence. Internal investigations are a part of the ethical infrastructure in organizations (Treviño et al. 2014). They fall under the scope of formal surveillance and sanctioning systems, which are standardized and visible procedures that monitor and detect behaviors and that associate behaviors with formal punishments and rewards (Tenbrunsel and Messick 1999, Tenbrunsel et al. 2003, Treviño et al. 2014).

As more organizations establish formal sanctioning systems, an increasingly important question is whether sanctioning systems are effective at influencing people's ethical decision making. Ideally, effective systems should be a source of moral and normative influence and be successful at not only preventing wrongdoing behaviors, but also at facilitating reporting and helping in the workplace (Institute of Business Ethics 2015). Unfortunately, findings on sanctioning systems' overall effectiveness are mixed (see Tenbrunsel and Smith-Crowe 2008, Treviño et al. 2014, for a review) and we still lack empirical understanding of how sanctioning systems influence different ethically relevant behaviors, and how they impact people's moral and normative psychological processes.

Existing research has mostly emphasized how sanctioning systems only trigger economic considerations while crowding out ethical motivations (e.g., Tenbrunsel and Messick 1999, Warren and Smith-Crowe 2008). However, some studies have also suggested that sanctions may hold moral meaning (e.g., Lindsay et al. 1996, Mulder et al. 2009) and signal social norms (Fehr and Fischbacher 2004). This leads us to our next point; most sanctioning systems research has primarily examined the punishment dimension, focusing on varying the presence (Gneezy and Rutschini 2006, Yamagishi 1986) or severity of punishment (Mulder et al. 2009). Surprisingly, much less research has been done on the investigative component of these sanctioning systems, even though the prospect of detection and getting caught is an important consideration in ethical decisions (Gellerman 1986, Nagin and Pogarsky

2003). Moreover, the majority of sanctioning system studies have adopted a single-party perspective, with a primary focus on actor behaviors (e.g., Hegarty and Sims 1978, McCusker and Carnevale 1995, Tenbrunsel and Messick 1999). However, many ethical decision-making situations in the workplace are complex multi-party *interactions* between *actors*, *recipients*, and *observers* who each have different behavioral options (see Brockner et al. 2015 for a review). One prominent example is interpersonal misconduct, which is the most common type of wrongdoing in the workplace (ERC 2013). Specifically, *actors* can commit harmful actions that impact the outcomes of themselves and of recipients. *Recipients* directly experience the actions of actors and how they respond influences their own outcomes and those of actors. *Observers* do not experience an actor's behavior first-hand (they hear about or witness the wrongdoing), but whether they help recipients is consequential to their own outcomes as well as to those of actors and recipients. Regardless of their roles, all parties are important contributors to event outcomes, and recipients and observers should be considered along with actors (Brockner et al. 2015).

Thus, the purpose of our research is two-fold. First, we aim to build a theoretical case for how investigative sanctioning systems influence people's moral and normative psychological processes and ethically relevant behaviors. Second, we seek to contribute to the scarcity of empirical work on the effectiveness of sanctioning systems (particularly the detection characteristic of these systems) in *multi-party behavioral interactions*. In contrast to most literature, we argue that sanctioning systems do not necessarily crowd out ethical concerns and we propose that they can also influence ethical and normative processes to affect people's ethically relevant behaviors. Theoretically, we build upon and extend Tenbrunsel and Messick's (1999) two-stage "signaling-processing" decision-making model and we draw from self-determination theory (Ryan and Deci 2000) to explain how and why sanctioning systems can affect people's moral and normative psychological processes and their ethically relevant behaviors. We suggest that stronger (versus weaker) investigative sanctioning systems can lead to (1) increased moral judgments that wrongdoing behaviors are unethical and (2) decreased social norm perceptions that wrongdoing behaviors are typical. In turn, these moral judgments and social norm

perceptions about wrongdoing lead to less undesirable and more desirable behaviors.[1]

We aim to make several contributions with our research. First, we add to the behavioral ethics literature by increasing understanding of sanctioning systems' effects on different ethically relevant behaviors within the same interaction. Departing from extant work that has taken a single-party perspective, we adopt and test a multi-party perspective to provide a more holistic understanding of the influence of sanctioning systems. Second, we extend prior work on sanctioning systems that has been largely dominated by a rationalist perspective, and we explain how sanctioning systems can also have moral and normative influences on behaviors. Third, we contribute to the behavioral ethics and justice fields by developing an interactive method to study both psychological processes and behavioral dynamics related to interpersonal harm in the lab. We designed a novel experimental game (see Weingast 1997 for a conceptually related game) to observe whether sanctioning systems influence people's ethically relevant psychological and behavioral reactions in multi-party wrongdoing interactions. If we want a more nuanced understanding of how organizational variables, such as sanctioning system strength, impact behaviors in socially interdependent situations with multiple focal parties, it is crucial to observe their impact in interactions that involve actual behavioral decisions with real consequences. Because experimental behavioral games are highly involving, such that participants care about what happens to themselves and to others, they can be parsimonious, precise, and coherent models that meaningfully capture otherwise complex social interactions (Camerer 2003, Murnighan and Wang 2016).

In addition to theoretical and empirical relevance, our research has organizational relevance. Accurate detection is a critical component of many companies' formal ethical infrastructures, and

[1] We distinguish *unethical* from ethical behavior based on Ashforth and Lange's (2016) broad definition of unethical actions as "self-interested behavior that intentionally or unintentionally harms others…in a way that generally would be considered unjustified and inappropriate by third-party observers," (p. 305). We define *normative* behavior as actions that are perceived as commonplace and typical (O'Leary-Kelly et al. 2000); thus, this definition mainly refers to *descriptive norms*. Supporting Ashforth and Lange's (2016) definition of unethical behavior, which contains prescriptive elements, Warren and Smith-Crowe (2008) argued that moral judgments are made of two key components: moral rules about how one "should" behave and moral beliefs about the state of the world. Thus, in our paper, moral judgment refers to judgments that can include moral rules and beliefs about behaviors whereas normative perceptions refer mainly to perceptions of what people typically do.

examining the influence of sanctioning systems on wrongdoing, reporting, and helping behaviors is particularly relevant for organizations today. Consider the 2017 "#MeToo" and "#TimesUp" social media movements about the magnitude of interpersonal harassment and mistreatment in workplaces. Some questions in the media about these movements center around the presence of due process and the accuracy of investigative findings, raising concerns about the possibility of false allegations and the assumption of guilt by accusation (e.g., see Atwood 2018, Stephens 2017, Williams 2018). These concerns about accurate detection when allegations of misbehavior arise highlight how an institution's capability to detect alleged wrongdoing can impact people's judgments, perceptions, and behavioral reactions to misbehaviors. To further highlight the organizational basis and relevance of our work, we also incorporate data from a supplementary textual analysis study of the codes of conduct of the largest 100 U.S. employers (see Online Appendix A for details) to inform our theory and hypotheses. Moreover, to ensure that our experimental game mirrored actor, recipient and observer interactions in interpersonal wrongdoing situations in organizations, its design was built on insights from a dissertation-based (Chui 2015) qualitative field study with professionals about workplace interpersonal misconduct (Chui 2016).

PERCEPTIONS OF SANCTIONING SYSTEMS AND THE DETECTION OF WRONGDOING

According to Lewin (1936), people often respond to their *perceptions* of situations rather than to an objective state of reality. Following this perspective, researchers have studied how sanctioning systems influence people's behaviors by manipulating people's subjective perceptions of sanctioning system strength. In extant literature, this has been characterized as perceptions of the presence or absence of a system (Gneezy and Rustichini 2000), a low or high sanction severity (Mulder et al. 2009), and a weak or strong probability of wrongdoing detection (Tenbrunsel and Messick 1999). Most relevant to our paper is Tenbrunsel and Messick's (1999) characterization of sanctioning system strength as the probability of wrongdoing being detected, in which they manipulated perceptions of system strength by varying the probability of wrongdoing detection (5% for weak; 50% for strong). Likewise, we operationalize perceptions of sanctioning system strength as the probability that wrongdoing will be detected.

In organizations, investigative sanctioning systems are typically operationalized as internal investigative committees that serve to uncover various employee wrongdoing ranging from interpersonal to financial misconduct.[2] Internal investigations can be conducted by Human Resources members or by a special committee of independent board members, and can vary in their ability to detect wrongdoing, depending on the existence of a reporting process, the expertise and experience of selected committee members, the independence and authority of the committee, the compensation of committee members, the ability to retain other experts, advisors, or legal counsel, and the breadth of the committee's rights to collect data (Markel and Murray 2017). Investigative activities can include interviewing key parties and accessing relevant technical and forensic data such as email review, cell phone records, videos, access-card logs, inventory audits, or financial records, among others (e.g., Glazer 2015, Slater 2017), and using machine-learning-based malfeasance detection systems (e.g., see Valkanas et al. 2018 for fraud and cyberbullying detection examples). As a result of these recent emerging technological advances, organizations are becoming ever more capable of being equipped with very strong investigative systems that are increasingly effective at accurately detecting wrongdoing.

Although other research has focused on perceptions of enforcement (e.g., Treviño and Weaver 2001) as a relevant characteristic of sanctioning systems, we focus on studying perceptions of detection. People value fairness (Folger 2001) and detection is essential to the perceived accuracy (and thus fairness) of an organization's sanctioning system. Conceptually, we view detection as an equally important process that should ideally occur prior to enforcement. Without detection as an initial step in the process, enforcement alone would breach a basic human right in most societies – the presumption of innocence or the assumption of being correct (e.g., United Nations Universal Declaration of Human Rights), which can be perceived as a form of procedural injustice (Cropanzano and Greenberg 1997). We

[2] Internal investigations are particularly useful for detecting covert wrongdoing. Given that the goal of most internal investigations is to address a wide range of interpersonal to financial misconduct, internal investigative systems are often intentionally ambiguous about their target of focus until a specific misbehavior is reported, in which case a special committee or task force is organized. Results from our codes of conduct study (Online Appendix A) show that most firms refer to both financial and interpersonal misconduct as possible targets of investigations.

therefore focus on the detection characteristic of sanctioning systems in our paper, and we assume that enforcement follows detection by pegging the enforcement of rewards and punishments to detection probability rates in our experimental design. We define the strength of investigative sanctioning systems as how good a system is at detecting wrongdoing (e.g., Tenbrunsel and Messick 1999), such that stronger (versus weaker) systems have a higher probability of correctly detecting if wrongdoing truly occurred.

The Traditional Paradigm: An Economic Perspective

Traditionally, researchers have relied on theories grounded in economic thinking when describing the effects of sanctioning systems on people's behaviors, drawing on reinforcement (Skinner 1953) and agency theories (Fama 1980, Jensen and Meckling 1976). According to these theories, sanctions and incentives are important means to control and direct behaviors because people naturally aim to maximize benefits and minimize costs to themselves, such that people commit actions because of the anticipation of valued outcomes (Becker 1968). In other words, the economic paradigm suggests that people will refrain from behaving in ways that are costly to themselves. For instance, if the probability of detection and punishment is high, actors will likely not commit wrongdoing because the expected benefits of the wrongdoing behavior are likely lower than the costs of getting caught. In contrast, recipients will be more likely to report an actor's wrongdoing behavior given the likelihood that potential reporting costs will be lower than the benefits (e.g., recipient compensation, actor punishment). Following this economic perspective, Tenbrunsel and Messick (1999) proposed in their two-stage signaling process model that a sanctioning system's effect on behavior is purely economic. Specifically, their model suggests that because sanctioning systems change the economic payoff structure of decision-making situations (in Stage 1), they only activate considerations about economic costs-benefits while crowding out ethical considerations about the morality of the wrongdoing behavior (in Stage 2).

In our research, we acknowledge the important role of economic considerations by making economic payoffs salient in our experimental game. However, we argue that solely explaining sanctioning systems' effects via an economic paradigm does not take into account the moral and normative attributes that sanctioning systems also signal (Fehr and Fischbacher 2004, Mulder et al. 2009), which are attributes

that decision-makers value in addition to economic ones (Gino et al., 2016). For instance, research on self-concept maintenance has shown that in addition to economic self-interest, people simultaneously care about maintaining a positive self-concept (Mažar et al., 2008) – that is, as a person who complies with internal moral and normative standards. Indeed, it has been argued that people do, in fact, care about behaving in accordance to the moral imperative of a situation (Jones 1991), such as acting fairly or reacting to injustice by restoring fairness (Folger 2001). Based on these arguments, we thus propose that sanctioning systems will also have parallel effects on people's moral and normative considerations. It is important to note that we do not intend to argue whether moral and normative processes play a more important role than an economic process. Our primary intent is to show that sanctioning systems do not exclusively influence economic motivations at the expense of moral and normative concerns, and we argue and show that sanctioning systems can also affect people's moral and normative processes to influence their behaviors. In the next section, we introduce a morally and normatively motivated perspective of sanctioning systems' effects.

AN ALTERNATIVE PERSPECTIVE: SANCTIONING SYSTEMS AS MORAL AND NORMATIVE SIGNALS

Past research has shown that environmental cues can activate perceptions about which behaviors are morally appropriate and normatively common in a given situation (e.g., Cialdini et al. 1990, Desai and Kouchaki 2015), particularly when they pertain to people's immediate goals (Aarts and Dijksterhuis 2003).[3] Sanctioning systems, as an environmental cue, can be strongly associated with moral and normative concepts in people's minds concerning which behaviors are considered right or wrong, and common or deviant (Cooter 1998). For instance, research shows that sanction severity is associated with stronger justice perceptions (Podsakoff et al., 2006), increased moral condemnation (Mulder et al. 2009), and increased enforcement of cooperative norms (e.g., see Fehr and Fischbacher 2004 for a review).

[3] Research has shown that direct observation of others is not necessary to activate norm perceptions – physical environment cues can also activate norm perceptions (Aarts and Dijksterhuis 2003) due to the automaticity of situational norms (Schank and Abelson 1977) that can be readily activated and accessed when required.

A societal example of the expressive power of formal sanctioning systems is the legal system. The legal literature has widely theorized that the law and its courts convey moral norms and social values (Dharmapala and McAdams 2003). Given that organizational sanctioning systems mirror the legal sanctioning system's objective to investigate and detect misconduct – and to discipline wrongdoers and compensate victims if so – we propose that sanctioning systems can likewise signal to members an organization's moral values and social norms concerning wrongdoing. In other words, the strength of a sanctioning system (i.e., its ability to uncover wrongdoing) can signal the extent to which the organization morally and normatively tolerates wrongdoing (and the extent to which it values preventing wrongdoing).

Strong systems with high detection rates signal that misbehaviors are considered unethical and atypical enough to warrant efforts towards their prevention. Thus, stronger (versus weaker) systems are associated with a higher intolerance for wrongdoing; wrongdoing is more often accurately detected and sanctioned, conveying an environment's stronger beliefs that wrongdoing is morally unacceptable and socially atypical. In contrast, weaker systems (or the absence of a system) can signal the opposite: that the organization tolerates wrongdoing and is less concerned about its prevention. Weak systems with infrequent detection and sanctions mean that people may commit wrongdoing for some time before being impeded, during which wrongdoing may already be positively reinforced in the environment as morally acceptable and socially common (O'Leary-Kelly et al. 2000). In the next section, we discuss the roles of moral judgment and social norm perceptions of wrongdoing on actor, recipient, and observer decisions.

Moral Judgment, Social Norm Perceptions, and Actor, Recipient, and Observer Behaviors

Regardless of roles (e.g., actor, recipient, observer), people have a fundamental need for moral values and social norms to be upheld (Folger 2001). People desire to act ethically, to be treated fairly, and to adhere to commonly accepted behavioral norms (Brockner et al. 2015). Previously, we explained how sanctioning systems signal an environment's moral and normative values (e.g., Mulder et al. 2009, Fehr and Fishbacher 2004). Here, we propose that the aforementioned ethical and normative motivations triggered by sanctioning systems play an important role in people's wrongdoing, reporting, and helping decisions.

It has been argued that the more that people judge a behavior or situation as morally intense, the more likely they will act in ways that respond to the moral imperative of the situation (Jones 1991). Indeed, moral judgment has been widely proposed in the ethics and justice literatures to predict a variety of ethically relevant behaviors. For instance, many ethical, whistleblowing, and third-party intervention theoretical models in the management literature feature moral judgment as an important behavioral antecedent (for actor ethical decision-making models, see Treviño 1986 and Rest 1986; for a recipient whistleblowing model, see Gundlach et al. 2003; for observer intervention models, see Bowes-Sperry and O'Leary-Kelly 2005, O'Reilly and Aquino 2011, and Skarlicki and Kulik 2004). Thus, we reason that a stronger (versus weaker) sanctioning system's effect on increased judgments that wrongdoing is unethical will lead to more actor wrongdoing and less recipient reporting and observer helping behaviors.

Similarly, considerable research has demonstrated the important role social norm perceptions play in predicting a wide variety of ethically relevant behaviors (e.g., see Fehr and Fischbacher 2004, Pierce and Snyder 2008). Because humans are social beings, people are heavily influenced by norm perceptions about what constitutes typical behavior in a given situation. Norms influence behaviors because they inform people of effective or adaptive behavior. Once activated, norms provide the knowledge needed to guide behaviors (Cialdini et al.1990). Because people have a fundamental need to feel related to others (Ryan and Deci 2000), they are motivated to behave in a socially expected way in their environments (Aarts and Dijksterhuis 2003). Even people who have yet to integrate group norms fully will act in ways consistent with perceived local norms (Leung and Tong 2003). Thus, we reason that a stronger (versus weaker) sanctioning system's effect on decreased perceptions that wrongdoing is typical will lead to less actor wrongdoing and more recipient reporting and observer helping. Below, we detail the moral and normative behavioral reactions that sanctioning systems evoke in actors, recipients, and observers.

Actors. The fact that people care about acting in accordance with perceived moral and social values (Folger 2001, Gino et al., 2016) speaks to the question of how sanctioning systems trigger morally and normatively motivated behavior in actors so that they engage in rule following behavior (e.g., not committing wrongdoing) when presented with a wrongdoing opportunity for personal gain. An actor's

wrongdoing motive is more likely to manifest itself under a weak system because a weak system is ambiguous about whether wrongdoing behavior is considered unethical and uncommon, which can trigger an actor's own judgment that the behavior in question is "not that bad" and "typical". This gives an actor more liberty to engage in selfish wrongdoing while keeping his or her positive self-concept as a person in accordance with internal moral values and social norms. In contrast, an actor is less likely to commit wrongdoing under a strong system because strong systems send a clear signal that wrongdoing not only comes with high economic costs, it is also a violation of moral and social standards. Given that people desire to be ethical and act normatively (Folger 2001, Gino et al. 2016), actors will be less motivated to commit wrongdoing under stronger systems and will be more likely to follow the system's regulations because of stronger moral and normative concerns that the behavior is considered unethical and deviant.

Recipients. A basic tenet in the justice literature is that people prefer to be treated fairly by others (e.g., see Brockner et al. 2015 for a review). This speaks to the question of how sanctioning systems affect morally and normatively motivated considerations in recipients so that they engage in rule utilization behavior (e.g., reporting wrongdoing) when reacting to experienced wrongdoing. Recent research has suggested that ethical and social norm motivations are determinants of fairness (e.g., Brockner et al. 2015). When a recipient's right to be treated fairly (i.e., ethically and according to social norms) is violated by an actor, recipients are at the onset, typically motivated to react and self-restore fairness (e.g., Brebels et al. 2008, Folger 2001), such as by reporting on the actor. However, weaker systems can discourage such responses because they can decrease unethicality judgments of an actor's behavior and increase norm perceptions that such behavior is common. This decreases a recipient's confidence that their experience is considered unjust and unethical, which can subdue their motivations to report. Indeed, it has been argued that even when recipients recognize a behavior as negative, they will be reluctant to challenge it if it is considered as common and normative behavior (Riger 1991). Conversely, stronger systems that signal and trigger increased unethicality judgments and decreased perceptions that wrongdoing is a behavioral norm emphasize in recipients' minds that an injustice has indeed occurred. Thus, stronger systems are more likely to increase recipient reporting, not only because it is more

economically viable to do so, but because stronger systems confirm recipient perceptions that they have been subjected to injustice, which can empower underlying recipient motivations to restore fairness.

Third-Party Observers. Observers play an important role in ensuring that wrongdoing gets detected and punished in organizations (Bowes-Sperry and O'Leary-Kelly 2005) given the passivity of recipient reactions (Knapp et al., 2007) and the ubiquity of observers (Skarlicki and Kulik 2004). That most people care about fairness as an end in itself (Folger 2001) speaks to why sanctioning systems motivate moral and normative behavioral reactions in observers so that they engage in rule supporting behavior (e.g., supporting recipient reports of wrongdoing) when confronted with wrongdoing. Unlike direct recipients, observers are not directly affected third-parties – thus, the economic paradigm suggests that observers will not help, if helping is costly. However, research has shown that observers *do* help, and that their decisions to do so are often morally motivated (Skarlicki and Kulick, 2004), despite economic costs to themselves (Fehr and Gächter 2002, Turillo et al. 2002). This suggests that observers may not only be motivated by a deterrence rationale but also by a just-desserts rationale (Carlsmith et al. 2002). Importantly, observers are more likely to help when they deem that an actor's behavior violates certain moral values and social norms (Skarlicki and Kulik 2004), because this confirms that an injustice has occurred. Thus, we reason that stronger systems that convey wrongdoing behavior as more unethical and less socially normal will increase observer helping because observers are more likely to attribute an injustice when they have stronger beliefs that moral and normative standards have been violated.

Taken as a whole, we thus propose that sanctioning system strength will influence morally and normatively motivated actor wrongdoing, recipient reporting, and observer helping behavioral decisions.

> **Hypothesis 1**: Stronger (versus weaker) sanctioning systems will lead to less actor wrongdoing (*Hypothesis 1a*), more recipient reporting (*Hypothesis 1b*) and more observer helping (*Hypothesis 1c*).

In the next section, we discuss how and why sanctioning systems shift people's moral judgments and social norm perceptions about wrongdoing to match those signaled by their environment's system.

Shifting Moral Judgments and Social Norm Perceptions about Wrongdoing

According to self-determination theory (SDT), most behaviors after childhood are extrinsically motivated (Deci and Ryan 1985, Ryan and Deci 2000). However, SDT proposes that extrinsic motivation can vary in its relative autonomy from less autonomous (i.e., external and introjected) to more autonomous (i.e., identification and integration), such that extrinsic motivation at the highest level of autonomy (i.e., integration) shares similar qualities with intrinsic motivation. Central to SDT is the idea that extrinsically motivated actions, such as actor wrongdoing, recipient reporting, and observer helping motivated by sanctioning systems, can become self-determined, as people identify with the environment's endorsed values, beliefs, and behaviors, and assimilate their regulation (Ryan and Deci 2000).

An important tenet of SDT (Ryan and Deci 2000) is that human beings have three universal and innate psychological needs that they try to satisfy and which are essential to their wellbeing – autonomy (to be the source of one's own behavior), competence (to feel effective), and relatedness (to feel connected to others). One way that people achieve and satisfy these needs is to internalize and integrate their external environment's endorsed values and behaviors (Ryan and Deci 2000). SDT proposes that people will express their autonomy, competence, and relatedness in different ways within external environments that endorse different values, such that the way people fulfill their needs satisfaction are influenced by the "ambient demands, obstacles, and affordances in their sociocultural contexts" (Ryan and Deci 2000, p. 75). Indeed, related research on cognitive dissonance (Festinger 1954), socialization (Campbell 1964), and positive self-concept maintenance (Greenwald 1980) lends support to the needs satisfaction motivation perspective. People desire to resolve cognitive dissonance (Gilovich et al. 2005), to assimilate the values and norms of their environment (Henrich et al. 2001), and to maintain a view of oneself as a person who complies with internal moral and normative standards (Greenwald 1980). One way to decrease ethical and social normative dissonance with their environment is to alter internal moral and social norm beliefs about a given wrongdoing situation to match those of their environment (Ashford and Anand 2003). Supporting this perspective, research has shown that moral judgments and social norm beliefs are indeed malleable and dependent on the environmental context (e.g., Gino 2015).

Given the fundamental role that needs satisfaction plays in people's behavioral decisions, we suggest that when individuals are presented with a wrongdoing situation (e.g., whether to commit wrongdoing or to respond to wrongdoing), they are likely to shift and integrate their internal moral and normative values and beliefs about the wrongdoing in question to match their sanctioning system's endorsed values and beliefs, even if their judgments and perceptions may be different in other environments (e.g., at home). For example, for individuals who may at the onset share similar internal moral and normative beliefs about wrongdoing as signaled by the sanctioning system, SDT suggests that their judgments and perceptions will likely be further strengthened to align closer to their sanctioning system. For individuals who may *not*, at the onset, share similar internal moral and normative beliefs about wrongdoing as endorsed by their sanctioning system, SDT suggests that they will likely shift and assimilate their moral and normative beliefs to move closer to their sanctioning system's, so that they will feel more self-determined in their subsequent behavioral motives (that are influenced by the sanctioning system that they are in). By identifying and integrating the wrongdoing in question as more unethical and less typical in stronger sanctioning systems – and as less unethical and more socially normal under weaker sanctioning systems – individuals are able to fulfill their fundamental needs for competence, autonomy, and relatedness in their behavioral decisions under different sanctioning system environments.

Consider actors who are tempted by an opportunity to commit self-interested wrongdoing. Actors are unlikely to do so in stronger systems because their needs for competence, relatedness, and autonomy are more likely undermined if they commit wrongdoing (e.g., it is ineffective because they will likely be caught, it disconnects actors from others because it goes against the stronger system's values and beliefs). Thus, to feel more self-determined about *not* committing wrongdoing under stronger sanctioning systems, actors will likely shift their moral judgments and social norm perceptions (i.e., that wrongdoing is more unethical and less socially normal) to align more closely to stronger systems. Conversely, actors are more likely to commit wrongdoing under weaker sanctioning systems. Weaker systems convey a higher tolerance for wrongdoing – thus, wrongdoing is less likely to threaten the needs satisfaction of actors with wrongdoing motives (e.g., wrongdoing is unlikely to be caught and is less disconnected from the group's

moral and normative behavioral standards whereas the low likelihood of detection and punishment allows actors to autonomously commit wrongdoing). Thus, actors' judgments and perceptions of wrongdoing (i.e., as less unethical and more typical) will likely be further strengthened under weaker systems.

Next, consider recipients and observers who have an opportunity to react to wrongdoing. It has been argued that both recipients and observers are typically motivated, at the onset, to restore fairness (e.g., Brockner et al., 2015). Recipients are more likely to report and observers are more likely to help in stronger systems because these behaviors are less likely to threaten their needs satisfaction (e.g., both reporting and helping behaviors are likely effective at restoring fairness and are in line with the environment's moral and normative values and beliefs). Thus, recipient and observer judgments and perceptions (i.e., that wrongdoing is unethical and non-typical behavior) will likely be further enhanced under stronger sanctioning systems. In contrast, recipients are less likely to report and observers are less likely to help in weaker sanctioning system environments because their needs satisfaction is more likely undermined (e.g., reporting and helping are unlikely to be effective at restoring justice and are less in line with the group's norms and tolerance for wrongdoing). Thus, recipients and observers in weaker systems will likely alter their moral judgments and social norm perceptions about the wrongdoing in question to fit with the system to feel more self-determined in their decisions *not* to report or *not* to help.

Importantly, SDT is not necessarily a stage theory; one can readily adopt or partially integrate new values, beliefs, and regulations at any point in time (Ryan and Deci 2000). Moreover, research has suggested that even if such regulations are not initially fully integrated, people tend to move towards integration of their external environment's regulations, over time (e.g., Chandler and Connell 1987). Thus, we further suggest that investigative sanctioning systems can have both immediate *signaling* effects on people's moral judgments and social norm perceptions, as well as trigger *learning* effects that occur over time. Below we elaborate on the conceptual and practical distinction of sanctioning systems' initial signaling and later learning effects.

Signaling and Learning Effects of Sanctioning Systems. After people are initially exposed to a sanctioning system and its moral and normative signals, they behave in direct response to the perceived

system and also observe how others behave in response to the system. In organizations, this initial signaling effect on people's judgments, perceptions, and behaviors can be triggered during specific moments, such as when newcomers join an organization and start the socialization process. For example, when newcomers attend orientation, they are commonly informed about how their organization investigates employee misconduct in their training classes (IBE 2015) or in their code of conduct handbooks that they may be asked to read prior to starting their jobs. Consistent with this scenario, results from our supplementary study reported in Online Appendix A show that 99 of the 100 largest employers in the U.S. mentioned the presence of internal investigatory procedures in their codes of conduct. Indeed, past research has shown that newcomers are especially susceptible to organizational influence regarding values, norms, and behaviors (e.g., Van Maanen and Schein 1979). Because the initial entry period is a particularly impressionable time for newcomers tasked with unlearning previous assumptions and processing new cues (Louis 1980), it is likely that their first exposure to their new organization's sanctioning system (e.g., during orientation) will directly influence their initial ethical judgments and norm perceptions regarding misconduct and affect their subsequent behaviors.

As time passes, however, people's judgments, perceptions, and behaviors regarding misconduct also become learned and strengthened via direct experiences and via observing others' behaviors over time. Through continual moral and normative signaling from the sanctioning system and via increased experience reacting to the system and interacting with others under the same sanctioning system environment, people's judgments about unethical behavior and their perceptions about what is accepted as ethical and normal behavior are strengthened. In particular, perceptions about whether wrongdoing is considered a social behavioral norm will have a high impact on actor, recipient, and observer behaviors through operant conditioning because personally experiencing and observing what behaviors are common and are permitted in the organization shapes our perceptions and subsequent behaviors (e.g., Morris et al. 2015). In organizations with weak sanctioning systems, for instance, gradual exposure to the environment's tolerance for wrongdoing behavior, along with the personal experience of one's past behaviors and of observing others' behaviors, can serve to unmoralize and normalize wrongdoing

(Ashforth and Anand 2003), if over time, people frequently experience a) actors committing wrongdoing without being detected and sanctioned, b) recipients not reporting an actor's wrongdoing, and c) third-party observers not helping recipients of alleged wrongdoing. According to social learning theory (Bandura 1977), people learn by observing others' behavior and its consequences. People perceive what constitutes behavioral norms via sanctioning systems because harmful behavior that is detected and punished indicates that a norm was violated whereas harmful behavior that is undetected and unsanctioned indicates otherwise (Bettenhausen and Murnighan 1985). Indeed, we expect that sanctioning systems will have a particularly strong learning effect as time passes and people have more experience to observe and learn behaviors. Thus, we propose that the ethical and normative considerations that sanctioning systems evoke can take place at both initial and later stages. Specifically, sanctioning systems have a *signaling effect* on moral judgments and norm perceptions at both initial and later stages, with later judgments and perceptions also the result of *learning effects*.

> **Hypothesis 2 (Signaling Effect):** Stronger (versus weaker) sanctioning systems will lead to increased judgments that wrongdoing behaviors are unethical (*Hypothesis 2a*) and decreased perceptions that wrongdoing behaviors are typical (*Hypothesis 2b*).
>
> **Hypothesis 3 (Learning Effect):** Over time, stronger (versus weaker) sanctioning systems will lead to increased perceptions that wrongdoing behaviors are unethical (*Hypothesis 3a*) and decreased perceptions that wrongdoing behaviors are typical (*Hypothesis 3b*).

Finally, our aforementioned theorizing regarding the ethical and normative processes triggered by sanctioning systems also leads us to two mediation hypotheses. If stronger sanctioning systems are associated with the extent to which people judge a behavior as more unethical and less typical, this will increase the likelihood that people will behave in accordance to the *moral imperative* of the situation (Jones 1991) and decrease the likelihood that people accept wrongdoing behavior as normal (Bettenhauser and Murnighan 1991). This means that for actors, increased judgment that their behavior is unethical and decreased perceptions that their behavior is socially typical will lower their wrongdoing behavioral motivations. For recipients and observers, stronger moral judgments that the actor's behavior

is unethical and weaker perceptions that the actor's behavior is a common social norm will increase the likelihood that recipients and observers will feel both morally compelled to respond and socially accepted to challenge the behavior, such that recipients will be more inclined to report and observers will be more motivated to intervene. We thus propose that sanctioning systems' effects on actor wrongdoing, recipient reporting, and observer helping are mediated by moral judgments and social norm perceptions of the wrongdoing situation in question.

Hypothesis 4 Judgments about the unethicality of wrongdoing behaviors will mediate the relationship between sanctioning systems and actor wrongdoing (*Hypothesis 4a*), recipient reporting (*Hypothesis 4b)*, and observer helping behaviors (*Hypothesis 4c)*.

Hypothesis 5: Perceptions about the typicality of wrongdoing behaviors will mediate the relationship between sanctioning systems and actor wrongdoing (*Hypothesis 5a*), recipient reporting (*Hypothesis 5c)*, and observer helping behaviors (*Hypothesis 5c).*

Let us briefly return to Tenbrunsel and Messick's (1999) signaling-processing model. They proposed that sanctioning systems act as an environmental cue in Stage 1 of decision-making and signal that the situation should be framed economically. In Stage 2, because the economic lens was adopted, the decision-making situation will be perceived as having primarily economic considerations. Here, we build upon their work and propose an alternative explanation for the impact of sanctioning systems on people's ethically relevant decision making. We propose that investigative sanctioning systems can send an implicit message of wrongdoing intolerance (or tolerance) to people in Stage 1 and make them perceive the situation as one that also has ethical and normative (in addition to economic) connotations. Thus, we contend that all of these considerations – moral, normative, and economic ones – have an influence on behavior. In Stage 2, with these processes operating in parallel, we propose that people will *alter* their moral and normative beliefs to fit the sanctioning system that they are in to fulfill their needs satisfaction to feel self-determined by assimilating the perceived values and beliefs of their environment (as signaled by the sanctioning system). In turn, these moral judgment and norm perceptions influence their behaviors.

RESEARCH OVERVIEW

Background and Empirical Design

We use an incentivized experimental game in Studies 1 and 2 in which multiple parties interact with each other to capture and simulate the following basic problem structure inherent in many workplace interpersonal wrongdoing situations: an individual (actor) can personally gain by harming another individual (recipient). The recipient can seek support and report to a third-party (observer) who is generally another member of the organization, and the observer makes a decision whether to help and support the recipient. The experimental game design of our studies was inspired by insights from qualitative interviews with business professionals about workplace interpersonal wrongdoing (Chui 2015, 2016). Transferring insights from employee personal accounts into an experimental game allowed us to study wrongdoing, reporting, and helping as a multi-party interaction with real behavioral consequences (e.g., Murnighan and Wang 2016). Below are key assumptions that we implemented in our game design.

Powerful Fixed Actor versus Alternating Recipient and Observer Roles. Most employees tend to alternate between roles as recipients and observers of wrongdoing whereas actors who mistreat others often have more power and authority than recipients and observers, such that actors mistreat recipients more often in the presence of less (vs. more) powerful observers (e.g., Chui 2015, 2016). Consistent with this finding, Brockner et al. (2015) proposed that organizational life makes it such that most employees indeed play multiple roles at work. For instance, people can be recipients who are mistreated by superiors and also observers who witness how their peers are being mistreated. Indeed, social power theory (French & Raven 1959) suggests that low power employees are more susceptible to power abuses and mistreatment, and research on workplace mistreatment and injustice (Cortina et al. 2001, O'Reilly and Aquino 2011) has argued that actors often have more power than recipients and observers.

The Importance of Observers. Supported by past work on observer reactions to mistreatment (e.g., Bowes-Sperry and O'Leary-Kelly 2005, Chui and Dietz 2014, Skarlicki and Kulik 2004), our design captures one example of wrongdoing involving actors, recipients, and observers, in which observers play an important role in supporting recipients and ensuring that wrongdoing gets detected and punished. Thus, our game is not generalizable to wrongdoing interactions in which observers may not

play a crucial role. We conceptualize observer decisions to support recipients as an analog to workplace investigations of certain harmful behaviors (e.g., interpersonal harm) that often require validation (e.g., corroboration, social support) to have a compelling case against the alleged before a formal investigation is launched (Mueller 2017, Murphy 2018). Because investigations are costly, firms are often advised to conduct case assessments in which case observer support can be an important determinant of whether a formal investigation is warranted (Jones 2012). Thus, observers in our design represent third-parties who can provide such support, corroboration, or validation for an investigation to be launched.[4]

Asymmetrical Actor Benefit and Recipient Harm. In our design, an actor's gain from wrongdoing is asymmetrical to a recipient's loss; recipients lose more than what actors gain from wrongdoing. This reflects many workplace wrongdoing situations in which recipients suffer more harm than what actors stand to gain. Consistent with this scenario, research on the effects of interpersonal wrongdoing has shown that recipients can suffer long-term harmful outcomes, such as emotional distress, withdrawal, decreased self-esteem and job satisfaction, and increased absenteeism (Cortina et al. 2017).

Actor and Recipient Reputation Loss. We included reputation costs for both actors and recipients whenever recipients reported. In organizations, mere accusations can affect reputations of actors and recipients (e.g., Atwood 2017), and recent research by Kennedy and Schweitzer (2018) has argued that whenever an accusation occurs, both the accused and the accuser suffer reputation losses.

Recipient Reporting and Observer Helping Costs. Supported by the literature in behavioral ethics and workplace injustice, we included reporting costs (e.g., Knapp et al. 1997, Treviño et al. 2001) and helping costs (e.g., Bowes-Sperry and O'Leary-Kelly 2005) in our game to acknowledge the costs that employees endure when deciding to seek justice.

---

[4] In our No Sanctioning System condition, we conceptualize observer helping as a type of vigilante behavior (e.g., O'Reilly and Aquino 2011). According to O'Reilly and Aquino (2011), in the absence of a sanctioning system, observers of injustice will personally dispense justice (e.g., firing the wrongdoer if high in power or ostracizing the wrongdoer if low in power), which can have a direct and consequential impact on actor and recipient outcomes.

Possibility of Recipient False Accusations and Observer Uncertainty. As argued in the media (e.g., Nelson 2015) and in research (Bies et al., 2015) accusations may not always be true, and our experimental game design takes this important element into account by allowing recipients the option of making false accusations and by introducing an element of uncertainty, such that observers are never 100% certain about the veracity of recipient allegations. Thus, our design mirrors the reality of imperfect knowledge when third-parties are exposed to recipient allegations of misconduct. Indeed, findings from Chui (2015, 2016) indicate that observers often felt uncertain about the credibility of wrongdoing allegations and that recipients cared about whether they might be believed by others. However, there is currently little to no research on false accusations in organizations, making it difficult to hypothesize how or why sanctioning systems might influence recipients to make false reports. Related research in criminal studies suggests that the motivations and circumstances behind false accusations are not necessarily the same as actively mistreating another via power exertion (Hall and Hall 2001). In sum, theory and empirical research on false accusation is currently limited (Kennedy and Schweitzer 2018), and the complexity of predicting recipient false accusations is further compounded by the multi-party behavioral interactions in our study. Therefore, although we allow for the possibility of false accusations in our game, we do not hypothesize for why or when recipients might make false reports under different sanctioning system environments. Instead, we explore and report recipient false accusations as an additional finding in Online Appendix D.

STUDY 1

Study Design

Figure 1 displays the decision trees and payoffs of our experimental game. The left panel displays the game in the sanctioning system conditions, with sanctioning systems of varying strength. The right panel displays the no system condition (see below for more details). The parameters on which payoffs are based are summarized at the bottom right of Figure 1, which we explain in more detail below.

[Insert Figure 1 about here]

Each interaction begins with every party (actor, recipient, observer) owning 100 points. If actors do not steal and recipients do not report, every party keeps their original endowment of 100 points (see the top-right end-node in Figure 1's game trees). If actors steal part of the recipient's points for personal gain, they decrease the recipient's points by 60 and increase their own points by 40. Stealing thus introduces overall inefficiency and is not a pure transfer, corresponding to the harmful effects of interpersonal wrongdoing in organizations.[5] After experiencing an actor's decision, a recipient can report on the actor to an observer. Importantly, like in real-life, recipients can also make false accusations. However, if recipients do not report actor stealing, the final outcomes (in points) are 140 (=100+40) for actors, 40 (=100-60) for recipients and 100 for observers (see top-left end-node in Figure 1's game-trees).

If recipients report, reporting costs recipients 15 points (e.g., mirroring costs such as reputation risk, job security) and actors 5 points (e.g., mirroring costs such as reputation loss). Importantly, recipient reporting gives observers the opportunity to help and support recipient reports. However, similar to real-life allegations in which there is always a possibility of false reports, observers are never 100% certain about the credibility of recipient reports. We modeled observer uncertainty using a binary but imperfect computer message indicating whether the actor stole; this message was only correct 80% of the time. Observers received this computer message before making their helping decision.

If observers help, helping costs observers 5 points (e.g., mirroring costs such as time) and actors an additional 5 points (e.g., additional reputation loss). Thus, if recipients truly report actor stealing, but observers do not help recipients, the final outcomes (in points) are 135 (=100+40-5) for actors, 25 (=100-

[5] To empirically assess whether participants perceived the decision to increase one's own points by 40 and to decrease another person's points by 60 to be unethical, we tested whether this behavior is considered to be inappropriate in the general population using Prolific Academic (N=100, 74% female, M_{age}=34.40, SD=10.36; see full results in Online Appendix B). We used six scenarios from Derfler-Rozin, Moore, & Staats (2016) who had adapted a set of questions from a dissertation-based field study about ethics from Martin (2016), and we created six additional scenarios for a total of 12 scenarios to establish in a sample of adults that Participant A's choice to increase his/her own points by taking away points from Participant B is perceived as 'inappropriate'. Results indicate that our focal behavior (M=2.95, SD=1.93) was perceived to be as inappropriate as underreporting one's taxes (M=2.61, SD=1.69, t(99)=1.42, p=.16), suggesting that the act of A taking away points from B in our study is perceived to be as unethical as committing tax fraud.

60-15) for recipients and 100 for observers. If recipients falsely report actor stealing and observers do not help, the final outcomes are 95 (=100-5) for actors, 85 (=100-15) for recipients, and 100 for observers.

We manipulated the strength of an investigative sanctioning system. Following Tenbrunsel & Messick (1999), we also included a no investigative sanctioning system condition for additional comparison. Study 1 thus had a total of four experimental conditions: weak system (55% detection probability) vs. strong system (85% detection probability) vs. very strong system (100% detection probability) vs. no system. System strength referred to the probability that an investigation, if launched, was able to correctly detect wrongdoing and dispense justice. Taking into account prospect theory, which also models how people perceive probabilities (Kahneman and Tversky 1979, Tversky and Kahneman 1992), we chose these probabilities to span the entire continuum from weak systems (a probability of 55% is only slightly better than chance) to fairly strong (85%) and eventually, in order to ensure that the detection probabilities are different enough in people's perceptions, to extremely strong systems (100%).[6]

In the sanctioning system conditions, if observers help recipients, an investigation (of varying strength) is launched. The investigation's outcome determines if an actor is sanctioned (by reducing his or her final payoff to 40 points), and a recipient is compensated (by receiving a compensation of 55 points). Thus, if actors steal and the investigation determines this correctly (which occurs with varying probabilities per the system strength), the final outcomes (in points) are 40 for the actor, 80 (=100-60-15+55) for the recipient, and 95 (=100-5) for the observer. False convictions are possible, however, and result in the same compensation and punishment. Thus, if an investigation incorrectly convicts an actor (who did not steal) for stealing, the final outcomes are 40 for the actor, 140 (=100-15+55) for the

[6]According to prospect theory's inversely s-shaped probability weighting function, the perceptual differences between probabilities in the middle of the spectrum are relatively small, such that people only perceive distinct differences between extreme probabilities (very low or very large). Thus, prospect theory suggests that two of our probabilities in the middle of the spectrum (55% vs. 85%) may subjectively be perceived as less different from each other than when the comparison involves an extreme probability (85% vs. 100%) even though the objective distance is larger in the first case. As such, we operationalize weak, strong, and very strong probabilities of detection as 55%, 85%, and 100% to ensure perceptual differences in participants. Importantly, the distinction between weaker and stronger systems is a qualitative distinction; the exact percentages and objective differences between 55%, 85%, and 100% systems are less relevant in this paper and it is the overall perception of strength and differences between perceptions of weaker and stronger systems that matter.

recipient, and 95 (=100-5) for the observer. In all true or false actor convictions, if investigations find that actors did not steal, the final outcome of the investigation does not trigger further monetary consequences (see final outcomes at the end-nodes of the "acquittal" branches in Figure 1's left panel game tree).

In the no sanctioning system condition, observer helping directly impacts final outcomes in lieu of a sanctioning system (e.g., see Tripp et al. 2007 for vigilante justice). If observers help, actors are punished and recipients are compensated according to the same logic described above (i.e., the actor's final outcome is 40 points, the recipient is compensated with 55 points). The final payoffs at the end-nodes of the "Support" branches in the game tree of the "no system" condition (see Figure 1) are therefore the same as those at the end-nodes of the "Acquittal" branches in the game with a sanctioning system.

Participants

We recruited 384 students (46.6% women, M_{age}=20.8 years, SD=2.4) from a paid university subject pool at large public university in Switzerland. Participants received 12 Swiss francs (the equivalent of 12 USD at the time of the study) as a show-up fee and could earn up to 31 Swiss francs more, depending on their and other participants' behavioral decisions during the study. Average payoff was 32 Swiss francs. We ran six sessions for each of the four conditions for a total of 24 sessions with 15 or 18 participants per session. In each session, participants engaged in 10 rounds of interactions with changing partners. We had 99 participants in the weak, 96 in the strong, and 93 in the very strong system conditions, and 96 in the no system condition.

Procedure

Participants were randomly assigned to a computer cubicle upon arrival and were given detailed written instructions describing the game. Participants had to pass control questions to ensure comprehension, and experimenters read aloud a summary of the instructions (see Online Appendix C2) to ensure common knowledge. No communication was allowed between participants. Participants were randomly assigned to roles A (actors), B (recipients), and C (observers). Participant As' (actor) role was fixed; they were always the ones who had the option to steal points. Participant B or Cs' (recipient or observer) role varied; they could alternate between being a B (recipient) or a C (observer) in each round, depending on

random assignment computerized on *z-tree*, an experimental software that allows for real-time anonymous interaction (Fischbacher 2007). The instructions and the experiment never used the words "actor", "recipient", and "observer" – rather, all parties were referred to neutrally as A, B, or C and all possible behaviors were described neutrally (see Online Appendix C1 for screenshots of the experimental software and how participants could make their decisions). All identities were anonymous, and everyone was assigned a pseudonym. Experimental conditions were randomly assigned at the session level.

Participants were informed that the experiment had two parts: the first part would consist of five rounds of the game, they would be paired with two others during each round, but they would not interact with any other participant more than once. After five rounds, experimenters implemented a surprise restart, announcing a second part of the experiment with another five rounds of the game. Everyone stayed in the same roles ("A" and "B or C") during the second part, but new pseudonyms were assigned. Importantly, participants did not know about the second part of five rounds when they interacted with each other in the first part, and at the start of the second part, they were truthfully told that there would not be any further part of the experiment involving a game. Our data therefore stem from a series of one-shot interactions because participants did not expect to meet their interaction partners again at any given point in time. We implemented multiple rounds to allow participants to become familiar with the investigative system and the decision environment, and to experience the consequences directly, allowing us to observe participants' behavior over time. Moreover, by collecting data from multiple interactions and decisions per participant, we ensured our results are not just artifacts due to initial confusion or a lack of experience (so-called "stationary replication", see, e.g., Camerer 1992 or Hertwig and Ortmann 2001).

Measures[7]

[7] In Online Appendix D, we provide analyses for our exploratory measures: false accusations (i.e., Bs reporting on As when As did not steal), and moral judgments and norm perceptions of stealing (these were collected after round 10, after all behavioral decisions). We extended the perception measures in Study 2 (see Study 2 for further details).

Behavioral decisions. The main dependent variables were the *percentages* of participant As choosing to steal, of participant Bs choosing to report (when As had stolen), and, finally, of participant Cs choosing to help Bs (conditional on B reporting on A to C and on a computer message matching B's report).

Results

Figure 2 displays the effects of different sanctioning system conditions on participant behaviors in Study 1 and shows the observed percentages for all behavioral decisions in each condition. Table 1 reports odds ratios from logistic panel regressions on each behavior,[8] with the weak system condition as the base category (see also pooled regressions and a figure displaying the development over time in Online Appendix E). Odds ratios below 1 indicate that the analyzed behavior (actor stealing in Column 1, recipient reporting in Column 2, and observer helping in Column 3) is less likely in the respective condition than in the weak system condition. Odds ratios greater than 1 indicate that it is more likely.

[Insert Figure 2 and Table 1 about here]

Actor Stealing. Actors stole the most in the weak system (68.8%) and the least in the very strong system (36.8%), whereas the stealing rates were similar in the strong (50.9%) and the no system (52.5%) conditions. Table 1 Column (1) shows regression results comparing stealing rates across conditions. Supporting Hypothesis 1a, actors stole significantly less in the very strong ($X_2(1)$=27.29, p<.01) and the strong system ($X_2(1)$=7.56, p<.01) than in the weak system. Actor stealing was not significantly different between the very strong and strong systems ($X_2(1)$=2.54, p=.11). Actors also stole significantly less in the very strong versus the no system condition ($X_2(1)$=8.52, p<.01), whereas the difference between strong system and no system condition was not significant ($X_2(1)$=0.63, p=.43). Interestingly, the weak system condition saw significantly higher levels of stealing than the no system condition ($X_2(1)$=8.08, p<.01).

Summary. Supporting Hypothesis 1a, stronger (vs. weaker) systems were better at decreasing actor stealing behavior.

[8] In the regressions, we corrected standard errors by clustering them at the session level. The reason is that participants interacted with each other and received feedback about the outcomes of these interactions after each round (see Cameron and Miller 2015, Fréchette 2012, and Moffatt 2015 for conceptual discussions and Bartling et al. 2017 or Kirchler et al. 2015 for applications in the analysis of multi-player-multi-round experimental games).

Recipient Reporting. Between the sanctioning system conditions, recipient reporting was highest in the very strong system (88.6%), followed by the strong (79.1%) and the weak systems (77.1%). In the no system condition, recipient reporting was even higher (93.5%). Table 1 Column (2) reports regression results. The difference in reporting between the very strong and the weak system was marginally significant ($X_2(1)$=2.92, p=.09). There were no further significant differences for recipient reporting between the sanctioning systems of varying strength (strong vs. weak: $X_2(1)$=1.80, p=.18; strong vs. very strong: $X_2(1)$=0.58, p=.44). The difference between the no system and the very strong system condition was also not significant ($X_2(1)$=2.13, p=.14). However, the no system condition was significantly better at promoting true reporting than the weak ($X_2(1)$=14.31, p<.01) and the strong system ($X_2(1)$=6.04, p=.01).

Summary. We found only marginal support for Hypothesis 1b, as the very strong system promoted marginally significantly more reporting than the weak. Interestingly, it was the no system condition that saw the highest levels of reporting (and significantly more than the weak and the strong system). However, as we show in Online Appendix D, the absence of a system also led to the largest number of false accusations, i.e. recipients reporting actor stealing when this was actually not the case.

Observer Helping. Observers helped the most in the very strong system condition (75.0%), followed by no system (70.3%), strong system (61.5%), and weak system (55.1%), when the computer message matched the recipient's report (i.e., also indicating that the actor had stolen points). Table 1 Column (3) reports regression results. Observers in the very strong system condition helped significantly more than observers in the weak ($X_2(1)$=3.88, p=.05) and marginally significantly more than observers in the strong system ($X_2(1)$=3.73, p=.05). The difference between the strong and the weak system was not significant ($X_2(1)$=0.17, p=.68). Although observer helping was directionally higher in the very strong system than in the no system condition, this difference was also not significant ($X_2(1)$=0.01, p=.91). However, compared to the no system condition, observer helping was significantly lower in the weak ($X_2(1)$=7.41, p<.01) and the strong system ($X_2(1)$=5.97, p=.01).

Summary. The data support Hypothesis 1c; sanctioning system strength had a clear impact on observer helping. The very strong system was significantly better at encouraging observer helping than

strong and weak systems. Interestingly, and in parallel to the finding on true reporting, we also find that observer helping was significantly lower in the weak and strong systems versus the no system condition, indicating that systems that are not strong enough can decrease observers' motivation to help.

Study 1 Summary of Behavioral Decisions. Overall, our data on actor stealing, recipient reporting, and observer helping support Hypothesis 1, such that stronger investigative sanctioning systems, compared to weaker systems, were indeed significantly better at discouraging actor stealing and promoting prosocial recipient reporting (marginal effect) and observer helping. Interestingly, the weak system led to worse results than the absence of a system for all behaviors such that there was significantly more stealing and less reporting and helping in the weak system versus the no system condition.

STUDY 2

We designed Study 2 to provide a more direct test of the signaling and learning effects triggered by sanctioning system strength (Hypotheses 2 and 3) and to shed more light on our proposed moral and normative mechanisms by testing whether they mediate the effects of sanctioning system strength on stealing, reporting, and helping (Hypotheses 4 and 5). We used the same experimental game paradigm as in Study 1; thus, Study 2 also serves as a replication test of Study 1's main results. However, there were two main differences. First, we extended the norm perceptions and moral judgment measures that we had included in an exploratory fashion in Study 1 (see Online Appendix D). Second, to be able to test both signaling and learning effects of investigative sanctioning systems, we measured the mediating variables (norm perceptions and moral judgment) *before* the first round (thus capturing the signaling effect) and *after* the final round of the game (thus capturing the learning effect).

To obtain a clearer picture of the difference between weaker and stronger systems and to simplify the study, we focused on examining the differences between two systems in Study 2: the weak (55% detection probability) and the very strong system (100% detection probability).

Participants

360 students (49.2% women, M_{age}=21.1 years, SD=3.2) from the same university subject pool as Study 1 (Study 1 participants were excluded from participating) received the equivalent of 10 Swiss francs (the

equivalent of 10 USD at the time of the study) as a show-up fee, and could earn up to 28 Swiss francs more, depending on their and other participants' behavioral decisions during the study. Average total payoff was 28 Swiss francs. We ran six experimental sessions for each of the two conditions for a total of 12 sessions with 30 participants per session. In each session, participants were randomly assigned to roles (A; B or C) and engaged in 10 rounds (two blocks of five rounds) of interactions with changing partners. Within each block, they would interact with each partner only once (like in Study 1).

Procedure

We used the same experimental procedures as Study 1 with two adaptations. First, as described above, we also measured moral and norm perceptions about wrongdoing before the first round, after participants read the instructions about the strength of the investigative sanctioning system. Second, due to additional laboratory space, we ran sessions with 30 participants each. To keep the size of the matching groups (within which participants are matched and re-matched with other players over the different rounds of the game) consistent with Study 1, we divided participants into two matching groups of 15 per session.

Measures

Behavioral decisions. Stealing, reporting, and helping decisions were the same as in Study 1.

Moral judgment and prescriptive norm of stealing.[9] Participants rated on a 5-point scale (1 = *strongly disagree* to 5 = *strongly agree*) the extent to which they thought stealing (i.e., A's decision to increase his/her points by decreasing B's points) was a) fair, b) unjust c) right, d) something one should

[9] To measure potential mediators, we extended the measures used in Study 1 (see Online Appendix D) to distinguish between prescriptive and descriptive norms. Thus, we aimed at measuring three mediating variables capturing moral judgment and norm perceptions of wrongdoing: moral judgment ("fair", "unjust", "right"), prescriptive norms ("something one should do", "not following social norms", "socially inacceptable") and descriptive norms ("normal", "majority does it", "common behavior", "typical behavior"). However, the measures for prescriptive norms (i.e., beliefs about what one ought or should do) and moral judgment (i.e., judgment about what is unethical) were highly correlated (correlation coefficients of .52 and .67, when measured before the first round and after the final round, respectively). To avoid multi-collinearity problems in the regression analyses, we merged the prescriptive norms and moral judgment measures for the analyses, thus focusing on two mediators capturing (i) moral judgment and prescriptive norms (i.e., the perceived unethicality and normative inappropriateness of stealing) and (ii) descriptive norms (i.e., the perceived typicality of stealing). There is also theoretical support for merging our moral judgment and prescriptive norm perceptions measures into one construct. As mentioned earlier, Warren and Smith-Crowe (2008) argued that moral judgments are comprised of moral beliefs about the state of the world and moral rules about what people should do.

do, e) not following social norms, and f) socially inacceptable, both *before* the first round (α_1=.75) and *after* the final round (α_2=.79) of the game. Items a), c), and d) were reverse-coded.

Descriptive norm of stealing. Participants indicated (both before the first and after the final round) on a 5-point scale the extent to which they thought a stealing decision by A was a) normal, b) something the majority would do, c) common, and d) something people would typically do (α_1=.78, α_2=.85). The measure thus captures how typical and common stealing is perceived to be.

Results for Actor, Recipient, and Observer Behavioral Decisions

Participant behaviors in Study 2 were very similar to the results from Study 1 with respect to the difference between the very strong and the weak system. Table 2 reports odds ratios from logistic panel regressions on each of the three behaviors, with the weak system condition as the base category.

[Insert Table 2 about here]

Actor Stealing. Consistent with Study 1, actors stole significantly less often in the very strong versus the weak system (35.5% vs. 70.2%, $X_2(1)$=18.18, p<.01).

Recipient Reporting. For recipient reporting (an effect that was only marginally significant in Study 1), the difference between the very strong and the weak system was of the expected positive direction but not statistically significant (79.8% vs. 75.5%, $X_2(1)$=1.40, p=.24).

Observer Helping. Similar to Study 1's findings, observers significantly helped more often in the very strong versus the weak system (65.2% vs. 48.1%, $X_2(1)$=6.39, p=.01).

Study 2 Summary of Behavioral Decisions. Study 2 replicated Study 1's effects of the very strong system (versus the weak system) for actor stealing and observer helping, supporting Hypotheses 1a and 1c. Importantly, the marginally significant effect for recipient reporting (Hypothesis 1b) from Study 1 did not replicate in Study 2 and was only directionally supported.

Results for Moral Judgment and Descriptive Norm Perceptions

[Insert Figure 3 about here]

Figure 3 displays the means of participants' perceptions (differentiated by participants' role in the experiment) about the unethicality and inappropriateness as well as the typicality of stealing when

measured *before the first* (left panel) and *after the final* (right panel) rounds of the experiment. Results show that the roles (A vs. B/C) to which participants were randomly assigned to had a main effect on moral and normative perceptions, such that participant As (actors) perceived stealing as less unethical and more typical than participant Bs/Cs (recipients/observers) in both the weak and very strong systems. This is consistent with the argument that actors adjust their moral and normative perceptions of wrongdoing in order to justify the act to themselves (see, e.g., Gino and Ariely 2012, Mažar et al. 2008). Importantly, however, the treatment effect of the very strong system (versus the weak) was the same regardless of role. Linear regressions of perceived unethicality and perceived typicality on a system dummy, a role dummy, and the interaction of the two yield non-significant interaction effects (see Tables E3 and E4 in Online Appendix E). For the following analyses, we therefore pool the data across the different roles.

Importantly, the results displayed in Figure 3 also indicate a signaling effect of sanctioning system strength (Hypothesis 2). Before the first round of the experiment, participants in the very strong system perceived stealing as more unethical and as less typical than in the weak system. There is also evidence of a learning effect (Hypothesis 3), as the differences in perceptions of wrongdoing between the weak and the very strong system are larger after the final round of the experiment than before the first round. We analyze these patterns, and their statistical significance, in more detail below.

Perceptions of stealing before the first round. As participants did not interact before answering the first set of perception measures, we use unclustered linear regressions (with heteroscedasticity robust standard errors) to investigate the statistical significance of signaling effects. We find that the very strong (versus weak) system led participants to perceive stealing as more unethical ($F(1, 358)$=4.87, p=.03, Cohen's d=0.23) and as less typical ($F(1, 358)$=5.86, p=.02, d=0.26). Thus, in line with Hypotheses 2a and 2b, we find a statistically significant signaling effect of the very strong system on both perception measures.

Perceptions of stealing after the final round. After the final round, we again measured participants' perceptions about the unethicality and the typicality of wrongdoing. Because measures were elicited after participant interactions during the experiment, we use linear regressions with standard errors

clustered by the 24 matching groups. Results show that the positive effects of the very strong system persisted and even grew stronger after participants had directly experienced the strength of the system over the ten experimental rounds. The very strong system (versus weak) led participants to judge stealing as more unethical and inappropriate ($F(1, 23)$=9.84, p<.01, d=0.31) and as less typical ($F(1, 23)$=25.01, p<.01, d=1.04). Thus, the very strong system also had a statistically significant effect on the measures of participants' moral judgment and descriptive norm perceptions after the final round of the experiment.

Changes in perceptions of stealing from first to final round. To investigate the learning effect triggered by sanctioning system strength further, we compare the measures for moral judgment and norm perceptions before the first and after the final round – and how their development differs between the very strong and the weak system condition. To do so, we use linear panel regressions, again clustering the standard errors by matching groups.

Within the very strong system, participants' perceptions shifted in a socially desirable direction – people judged stealing to be more unethical in the final round than in the first round ($X_2(1)$=14.34, p<.01). Moreover, in the very strong system, stealing was perceived as being significantly less typical at the end of the experiment than at the beginning ($X_2(1)$=21.67, p<.01).

Within the weak system, the perceived unethicality of stealing also increased significantly from the first to the final round ($X_2(1)$=5.54, p=.02). However, there was no difference between first and final rounds in the perceived typicality of stealing ($X_2(1)$=0.46, p=.50), In fact, mean perceptions of typicality in the weak system even directionally increased in the final round compared to the first round. When comparing the changes in perceptions about wrongdoing from first to the final round *between* systems, we find a significant difference between the weak and very strong systems for the perceived typicality of stealing ($X_2(1)$=18.12, p<.01), which decreased in the very strong system but (slightly) increased in the weak system. This is in line with Hypothesis 3a. For the perceived unethicality of stealing, the difference between the systems went in the expected direction, with the very strong system leading to a stronger change in perceived unethicality, but without being statistically significant ($X_2(1)$=1.50, p=.22). Thus, the evidence is only directionally in line with Hypothesis 3b.

Summary. The findings for moral judgments and norm perceptions are in line with our signaling effect hypotheses. The very strong system led participants to perceive stealing as more unethical (Hypothesis 2a) and less typical (Hypothesis 2b) even before the first round of interactions. The learning effect hypotheses were partially supported. The difference in perceived typicality of stealing between the very strong and the weak system became significantly stronger over time (in line with Hypothesis 3b) whereas the difference between the two systems for perceived unethicality (Hypothesis 3a) also increased over time but non-significantly. Thus, these results indicate that the learning effect triggered by the very strong system mainly occurs via descriptive norm perceptions about the typicality of wrongdoing.

Results of Mediation Analyses

To test whether perceived unethicality and perceived typicality of wrongdoing mediated the effects of sanctioning system strength on wrongdoing, reporting and helping (Hypotheses 4 and 5), we conducted multiple mediation analyses (Preacher and Hayes 2008) using PROCESS (Hayes 2013) with a logistic regression for modeling the behavioral decisions and relying on bias-corrected 95% confidence intervals (BCCI) for estimating indirect effects. Specifically, for each behavioral dependent variable (stealing, reporting, and helping), we conducted a multiple mediation analysis simultaneously entering moral judgment and norm perceptions about stealing as potential mediators of the effect of system strength.

In our mediation analyses, we further distinguish between the signaling and learning effects of investigative sanctioning system strength. In a first set of analyses, we test whether perceptions about stealing measured *before* the first round mediated the effects of system strength on first round decisions (to capture the signaling effect). In the second set of analyses, we do the same for final round decisions using the perceptions measured *after* the final round as mediators, thus capturing mediation after participants learned about the strength of the sanctioning system through direct and repeated experience.

In line with Hypotheses 4a and 5a, we find that the indirect effects of system strength on first round actor stealing decisions via perceived unethicality and perceived typicality were statistically significant: (via perceived unethicality: 95% BCCI = [-0.78, -0.01], via perceived typicality: 95% BCCI = [-1.19, -0.11]). For the other first round behavioral decisions (i.e., recipient reporting and observer

helping), the indirect effects were not statistically significant (Table E5 in Online Appendix E shows the full results for all multiple mediation analyses) and Hypotheses 4b-c and 5b-c are thus not supported.[10] For mediation effects of final round decisions, we again find support for Hypothesis 5a, such that the indirect effect of the very strong sanctioning system on stealing via perceptions of stealing as typical was highly significant (95% *BCCI* = [-2.00, -0.59]). In contrast to first-round decisions, and inconsistent with Hypothesis 4a, the indirect effect via moral judgment was not significant anymore at the 5% level in the analysis of final round decisions (95% *BCCI* = [-0.45, 0.06]). Again, we find no support for Hypotheses 4b-c and 5b-c, as indirect effects on final round reporting and helping decisions were non-significant.

Study 2 Summary of Mediation Analyses. The mediation analyses based on first round decisions support Hypotheses 4a and 5a. The very strong sanctioning system influenced actor stealing through an immediate signaling effect that made wrongdoing seem less typical and more unethical, with the route via typicality seemingly stronger and more robust. Similarly, the mediation analyses for the final round decisions further indicate that the learning effect triggered by directly experiencing the strength of the sanctioning system mainly worked via making stealing seem less typical (in line with Hypothesis 5a) and not via making stealing seem more unethical (thus not in line with Hypothesis 4a). Importantly, we only find significant mediation effects for actor stealing and not for recipient reporting or observer helping.

DISCUSSION

In this paper, we studied the effects of investigative sanctioning systems from a multi-party perspective. Specifically, we considered how investigative sanctioning systems affect different focal parties' ethically relevant psychological processes and behaviors. Our results show that in addition to decreasing wrongdoing in actors, the strength of investigative sanctioning systems is crucial for promoting prosocial behaviors of involved observers that are essential for preventing misconduct in organizations (e.g.,

[10] To check the robustness of our results, we also conducted simple mediation analyses, testing only one mediator at a time (see Table E6 in Online Appendix E). Compared to the multiple mediation analyses, the only difference in terms of statistical significance of indirect effects is that the 95% BCCI of the indirect effect of system strength via perceived unethicality on first round actor stealing overlaps with zero and is thus not statistically significant in the simple mediation analysis. We take this as evidence that the channel via perceived typicality is a more robust mediating mechanism of the effect of system strength on stealing than the channel via perceived unethicality.

Skarlicki and Kulik 2004). Moreover, we find that investigative sanctioning system strength affects actor, recipient, and observer moral judgments and social norm perceptions, such that very strong systems favor the perceptions that a given wrongdoing is more unethical and less common. These novel insights add to the widely acknowledged effect of sanctioning systems on economic incentives.

Theoretical and Practical Implications

Our paper makes a number of theoretical and practical contributions. First, we contribute to the behavioral ethics literature by presenting a more holistic understanding of sanctioning systems, such that wrongdoing is not examined as an isolated act, but rather occurs in the presence of others who will be hurt by it (i.e., recipients) and who can potentially intervene (i.e., observers), allowing for new theoretical perspectives. Across two experimental game studies, we find that stronger (versus weaker) investigative sanctioning systems have significant positive effects on actor wrongdoing and observer helping, and to an extent, a possible (marginally) positive effect on recipient reporting (observed in Study 1 only). Thus, adequately stronger systems successfully discourage actor wrongdoing.

Whereas less actor wrongdoing seems to be a natural consequence of stronger sanctioning systems, within the strategic interaction modeled in our study, this is only true if recipients are also motivated to report and observers are also motivated to help. Hence, the second main result of our study is equally relevant; if an investigative sanctioning system is too weak, it can negatively affect prosocial behaviors compared to when no system is present. Importantly, the very strong system condition avoided this negative effect on prosocial behaviors because increasing system strength significantly improved observer helping rates. In particular, even when moving from the strong to the very strong sanctioning system condition, this trend of system strength on prosocial behaviors remained. When comparing the strong (85% detection) versus the very strong (100% detection) system conditions in Study 1, observer helping was marginally higher in the very strong system than in the strong system. However, this distinction between the strong and very strong systems did not appear to matter for actor wrongdoing decisions, such that the very strong system was not significantly better at curbing actor wrongdoing than the strong system. Thus, our findings on the difference between strong versus very strong sanctioning

systems suggest the importance of recognizing the holistic effects of sanctioning systems, and of considering a variety of ethically relevant behaviors when deciding on an effective threshold of sanctioning system strength. As our results imply, some ethically relevant behaviors (e.g., observer helping) may be more sensitive to the difference between very strong systems and strong systems, than other behaviors (e.g., actor wrongdoing).

One possible explanation driving these results may be due to differences in the underlying behavioral motivations of different focal parties. For instance, actors with selfish wrongdoing motives may still care about acting in accordance with perceived moral and social values (Folger 2001, Gino et al. 2016). As such, a sanctioning system's strength only needs to reach a relatively "strong" threshold that effectively signals wrongdoing as more unethical and less normal (as well as more likely to carry a punishment risk), to already be sufficient at preventing wrongdoing, so much so that a very strong system may not make a significant difference in comparison. Observers, on the other hand, care about ensuring justice (O'Reilly and Aquino 2011, Turillo et al. 2008). This means that although they may desire to help actual recipients of wrongdoing, they are also reluctant to help potential false accusers (Skarlicki and Kulik 2004), even if such a possibility is small. However, because observers often rely on the interpretations of others (e.g., recipients), there is always an element of uncertainty when they come across wrongdoing allegations (Brockner et al. 2015). Thus, when observers are less certain about the veracity of wrongdoing allegations (and thus less confident that their intervention will actually help restore justice and not create injustice), they will be less likely to help. As sanctioning systems become stronger, however, this uncertainty is reduced, in part due to expectations that allegations are more likely to be true as well as higher confidence that even if false allegations arise, the investigative system will accurately uncover them as such. Because observer uncertainty was the most alleviated in the very strong system, even the difference between a strong versus a very strong system mattered.[11] As such, observers

[11] This reasoning is in line with our exploratory results (see Online Appendix D), in which recipient false accusations were indeed significantly lower in the very strong system than in the strong system (and this effect was very large), as well as compared to the weak and the no system conditions.

are likely to help less in a strong (versus a very strong) sanctioning system because a strong system still comes with a little more uncertainty than a very strong system, which can make observers less confident that their efforts will ensure justice. Importantly, the increase in observer helping as systems move from weak to very strong cannot be due to economic considerations because helping was equally costly in all conditions, and each round of interaction was one-shot and anonymous. Thus, our results show that observer helping motivations had to be stronger than self-interest, and that sanctioning systems do not only affect people's economic interests.

This leads us to our next contribution: our finding that sanctioning systems can significantly influence people's moral and normative psychological processes. By showing that investigative sanctioning systems have desirable effects on moral judgments and norm perceptions, we expand upon previous work (e.g., Tenbrunsel and Messick 1999) and offer an additional explanation for sanctioning systems' influence on ethically relevant behaviors. Our results show that economic considerations are not the only important signaling effects that investigative sanctioning systems have; as systems become stronger, they also signal attention to the ethical and social normative connotations of wrongdoing. The *signaling* effect of system strength in Study 2 is particularly supportive of this idea. Merely being informed that the system will be very strong (versus weak) made people perceive wrongdoing as more unethical and less typical. Moreover, sanctioning systems also affect moral judgment and norm perception via a *learning* effect. In strong systems, the experience that wrongdoing is reliably being punished not only decreases wrongdoing but also affects moral judgments and normative perceptions about wrongdoing over time. In particular, descriptive norms about wrongdoing are affected in a more desirable way over time in stronger (versus weaker) systems. Our mediation results show that both moral judgment and descriptive norms mediate the effects of system strength on initial wrongdoing behavior. However, they also show descriptive norms as being particularly important in the long run, as it is only descriptive norms that significantly mediate the effect of system strength on final round wrongdoing decisions in Study 2.

Additionally, by experimentally testing the influence of sanctioning systems on observer helping, we make an empirical contribution to the third-party observer intervention literature. Although there is a growing body of research on observer responses to injustice (e.g., see Skarlicki et al. 2015 for a review), most empirical studies have focused on how individual-level variables such as emotions (e.g., Mitchell et al. 2015) or situational-level variables such as type of injustice committed (e.g., O'Reilly et al. 2016) influence observer reactions. In particular, research on the environmental context and observer reactions has so far been theoretical (e.g., see Skarlicki and Kulik 2004, O'Reilly and Aquino 2011). Our work thus advances empirical knowledge on how environmental contexts affect observer processes and behaviors.

Finally, we demonstrate that investigative sanctioning systems do not influence all focal parties' behaviors in the same desirable way, highlighting again the importance of studying the overall effects of such systems. For instance, we demonstrate that although a weak system was less effective than the absence of a system for decreasing actor wrongdoing and for increasing recipient true reporting and observer helping, the absence of a system was also the most conducive for recipient false accusations (an exploratory finding reported in Online Appendix D). Only very strong systems were successful at simultaneously decreasing actor wrongdoing and recipient false accusations and at increasing observer helping. Thus, our research brings to light the importance of acknowledging the interdependence of interpersonal wrongdoing interactions, and we contribute to the behavioral ethics and justice literatures by developing a novel experimental design to study the multi-party social dynamics of interpersonal harm.

Of practical relevance, our results indicate that formal investigative sanctioning systems can be effective tools for deterring undesirable wrongdoing and for fostering desirable reporting and helping behaviors in organizations, which are issues that have recently gained increasing prominence in organizations (e.g., Bennett 2017). Importantly, we show that investigative sanctioning systems are not only effective by changing economic incentives, but that they also shape behavioral norms and ethical perceptions. The key insight from our research is that whether these systems effectively prevent wrongdoing and positively affect ethical and norm perceptions very much depends on their strength. Weak systems with a low probability of detecting and punishing wrongdoing can have negative rather

than positive consequences; they can send a cynical message to employees that systems are merely "window dressing" to protect organizations from potential liability (Treviño and Weaver 2001) rather than send a serious message to employees that the organization truly cares about preventing wrongdoing (with an effective system to prove it). Thus, in order to reap the full benefits of investigative sanctioning systems and to avoid counterproductive effects, it is important to make these systems strong enough.

Relatedly, the results of our no system condition in Study 1 revealed some interesting findings that are worth discussing further. Specifically, we find that a weak system was worse at decreasing wrongdoing and increasing prosocial reporting and helping than the absence of a system. There are at least two possible reasons for these counterproductive effects of weak systems. First, in the no system condition, people may be more likely to take matters into their own hands, because there is no system on which they can rely on to address experienced injustice (e.g., DeCelles and Aquino in press, O'Reilly and Aquino 2011). Thus, the no system condition encourages "vigilante justice", with recipients and observers recognizing the need to defend themselves and to report or help recipients of wrongdoing. Second, weak systems (versus stronger and no systems) may decrease motivations to report wrongdoing or help recipients of wrongdoing, especially over time, as participants learn that the (costly) acts of reporting and helping will often not lead to the desired results. In view of these results, some might argue that because certain types of misconduct may be difficult to detect even for very strong systems, the absence of a system might be better. However, similar to society's legal justice system (that may not catch every case of wrongdoing), having a strong system in place is arguably still better at preventing overall wrongdoing than the absence of a system, because it is the overall performance of a system's ability to detect and sanction a variety of wrongdoing behaviors (including false accusations) that matter.

Limitations and Future Directions

Despite the strengths of our research and findings, our work has several boundary conditions and limitations that need to be taken into consideration. First, our experimental studies were based on abstract laboratory game experiments and our operationalization of investigative sanctioning system strength was a probability percentage that wrongdoing would be detected by an investigative system. This allowed us

to study a complex multi-party phenomenon in a tractable way and to observe behavioral decisions with real consequences (see, e.g., Murnighan and Wang 2016). However, wrongdoing interactions in real-life settings are more complex and people's perceptions of an investigative sanctioning system's ability to detect wrongdoing in real organizations may hinge on various factors that may be less precise than the way participants perceived investigative sanctioning systems in our studies. Moreover, we deliberately chose extreme probabilities to represent weak and very strong sanctioning systems in our studies (i.e., 55% vs. 100% detection) to ensure that our manipulation was effective on participants' perceptions of system strength, due to past evidence (e.g., Tversky and Kahneman 1992) that people only perceive distinct differences between extreme probabilities. Although organizations are moving towards achieving near perfect detection capabilities due to emerging technological advances in organizational detection systems (Glazer 2015, Valkanas et al., 2018), we acknowledge that a literal 100% detection rate is currently unrealistic for many organizations. However, applying prospect theory (Kahneman and Tversky 1979), we contend that the exact percentages and literal differences between 55%, 85%, and 100% systems in our studies are less important; rather it is people's overall qualitative interpretations of the strength and differences between the weaker and stronger systems that matter. Future work should test the implications of our paper in field-experimental settings to provide further external validity to our findings.

Second, the role of observers in our research can be a limitation because our findings are specific to interpersonal wrongdoing situations involving observers. In our studies, observers play a crucial role by helping recipients ensure that actor wrongdoing gets detected and sanctioned (e.g., Skarlicki and Kulik 2004, O'Reilly and Aquino 2011). Because our research design involves behavioral interactions between actors, recipients, and observers within a specific setup (i.e., *both* recipient and observer behavioral decisions are crucial to whether an actor gets investigated and detected), an important boundary condition of our findings is that they may be less generalizable to other possible scenarios of interpersonal wrongdoing interactions in which observers may not play a role, or may play a less important role.

The interactive nature of our experiment also allowed us to consider false accusations as an exploratory finding (see Online Appendix D). We find that false accusations are most likely to occur in

the absence (versus the presence) of any investigative sanctioning system (in Study 1) and are more likely in weak (versus very strong) systems (in Study 2). One possible reason for these results could be that environments where wrongdoing and false accusations are unlikely to be accurately detected and sanctioned are seen as more conducive for those with false reporting motives. In such cases, false reports are more likely to manifest because those with false accusation motivations can merely make an accusation to bring upon the desired goal of inflicting harm on the accused without the risk of having to fulfill any formal burden of proof. Thus, although not having a system appears better than a weak system for preventing wrongdoing and encouraging reporting and helping, our exploratory findings on false accusations find that no systems can also be problematic due to a higher potential for false accusations. Surprisingly, research on false allegations in organizations has so far been limited, even though recipient credibility is an important consideration, leading to recent calls for more research on this critical topic (e.g., Kennedy and Schweitzer 2018). Our study thus makes a first empirical advancement in this area. Future studies should examine other antecedents and consequences of false accusations in organizations.

Finally, although we find that investigative sanctioning systems influence people's moral and normative perceptions about wrongdoing, which in turn, influence actor wrongdoing, we did not find significant mediation effects for recipient reporting and observer helping. Thus, it appears that other mechanisms may be driving reporting and helping. One possible reason why our proposed mediating mechanisms were not significant for reporting and helping is that our moral judgment and social norm perception items were focused on another's behavior (i.e., the actor's) rather than on one's own behaviors (i.e., reporting or helping). Although beliefs about an actor's wrongdoing behavior are important for assessing an interpersonal wrongdoing situation, judgments and perceptions about one's own behavior (versus another's behavior) may be more influential for motivating one's own (often costly) reactions to wrongdoing. As such, it seems that other moral and normative constructs that directly pertain to recipient reporting and observer helping (e.g., increased beliefs that reporting or helping is moral and normative) may be relevant mechanisms to consider in future studies. It is important to note, however, that although we found a marginal effect of stronger systems on recipient true reporting in Study 1, we did not replicate

this in Study 2. This suggests that our observed effect on recipient reporting is not conclusive. One possibility for this is that other important factors (e.g., negative emotions evoked by actor wrongdoing) may be triggering recipient reporting, which should be examined by future research.

There are other relevant questions that future studies should address. We mainly examined one type of wrongdoing (stealing something of value from another), and future work should examine other wrongdoing types (e.g., cheating). Finally, we focused on detection ability, and while detection is an important characteristic of a sanctioning system's quality (e.g., Becker 1968; Gellerman 1986), the nature of sanctions and enforcement rates could also have relevant effects. Future research should consider other aspects of sanctioning systems on wrongdoing, reporting, and helping behaviors.

CONCLUSION

Many organizations have established investigative sanctioning systems to detect and prevent unethical behavior. Previous research has mainly considered how such systems work by influencing economic motivations of first party actors and has cast some doubt on their psychological effectiveness by finding that investigative sanctioning systems may crowd out ethical concerns. Building upon this previous research, our research offers an alternative perspective and breaks new ground by studying the effects of investigative sanctioning systems from a multi-party perspective and by demonstrating that strong and effective systems that can detect wrongdoing and dispense justice with a high enough probability will also have desirable moral and normative psychological effects in addition to their positive behavioral effects. Stronger investigative sanctioning systems have a positive influence on moral judgments and descriptive norm perceptions about wrongdoing, and they decrease wrongdoing and increase helping behaviors. Thus, they can be important sources of not only economic influence but also of ethical and normative influence to effectively reduce unethical behaviors and promote prosocial behaviors in organizations.

REFERENCES

Aarts H, Dijksterhuis A (2003). The silence of the library: environment, situational norm, and social behavior. *Journal of Personality and Social Psychology*, 84(1): 18-28.

Ashforth B, Lange D (2016) Beware of organizational saints: How a moral self-concept may foster immoral behavior. Palmer, D, Greenwood, R, Smith-Crowe, K, eds. *Organizational Wrongdoing: Key Perspectives and New Directions* (Cambridge University Press), 305 – 336.

Ashforth BE, Anand, V (2003) The normalization of corruption in organizations. *Res. Organ. Behav.*, 25: 1-52.

Atwood M (2018) Am I a bad feminist?. *The Globe and Mail* (January 15), https://www.theglobeandmail.com/opinion/am-i-a-bad-feminist/article37591823/

Bandura A (1977) Self-efficacy: Toward a unifying theory of behavioral change. *The Psychological Review*, 84(2): 191-215.

Bartling B, Grieder M, Zehnder C (2017) Competitive pricing reduces wasteful counterproductive behaviors. *J. of Public Economics*, 156, 34-47.

Becker GS (1968) Crime and punishment: An economic approach. *J. of Political Economy*, 76: 169-217.

Bennett J (2017), The #metoo movement: When the blinders come off, *The New York Times* (November 30), https://www.nytimes.com/2017/11/30/us/the-metoo-moment.html

Bettenhausen KL, Murnighan JK (1985) The emergence of norms in competitive decision-making groups. *Admini. Sci. Quart.*, 30: 350-372.

Bies, R. J., Barclay, L. J., Tripp, T. M., & Aquino, K. (2016). A systems perspective on forgiveness in organizations. *The Academy of Management Annals*, 10: 245-318.

Bowes-Sperry L, O'Leary-Kelly AM (2005) To act or not to act: The dilemma faced by sexual harassment observers. *Academy of Management Review*, 30: 288-306.

Brockner J, Wiesenfeld B, Siegel P, Bobocel DR, Liu Z (2015) Riding the Fifth Wave: Organizational Justice as Dependent Variable. *Res. Organ. Behav.*, 35: 103-121.

Camerer C (1992) The rationality of prices and volume in experimental markets. *Organ. Behav. and Human Decision Processes*, 51: 237-272.

Camerer, C (2003) *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Cameron AC, Miller, DL (2015) A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50: 317-372.

Campbell EQ (1964) The internalization of moral norms. *Sociometry*, 27: 391-412.

Carlsmith KM, Darley JM, Robinson PH (2002) Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2): 284-299.

Cialdini RB, Reno RR, Kallgren CA (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *J. of Person. and Soc. Psych.*, 58(6): 1015-1026.

Chandler CL, Connell, JP (1987) Children's intrinsic, extrinsic and internalized motivation: A developmental study of children's reasons for liked and disliked behaviours. *British Journal of Developmental Psychology*, 5: 357-365.

Chui C (2015 July). *Observer reactions to workplace mistreatment.* (Unpublished doctoral dissertation). University of Lausanne, Lausanne, Switzerland.

Chui C (2016 April). *Observer reactions to workplace mistreatment: A qualitative study*. Poster session at the Society of Industrial and Organizational Psychology conference, Anaheim, United States.

Chui C, Dietz J (2014) Observing workplace incivility towards women: The roles of target reactions, actor motives, and actor-target relationships. *Sex Roles, 71*: 95-108.

Cooter R (1998) Expressive law and economics. *The Journal of Legal Studies*, 27(S2): 585-607.

Cortina LM, Magley VJ, Williams JH, Langhout RD (2001) Incivility in the workplace: Incidence and impact. *Journal of Occupational Health Psychology*, 6(1): 64-80.

Cortina LM, Kabat-Farr D, Magley VJ, Nelson K (2017) Researching rudeness: The past, present, and

future of the science of incivility. *Journal of Occupational Health Psychology*, *22*, 299-313.
Cropanzano R, Greenberg J (1997). Progress in organizational justice: Tunneling through the maze. *International Review of Industrial and Organizational Psychology*, 12: 317-372.
DeCelles, KA, Aquino, K. (In press). Dark knights: When and why an employee becomes a workplace vigilante. *Academy of Management Review*.
Deci E, Ryan RM (1985) *Intrinsic motivation and self-determination in human behavior* (Springer Science & Business Media).
Derfler-Rozin, R, Moore, C, Staats, BR (2016) Reducing organizational rule breaking through task variety: How task design supports deliberative thinking. *Organization Science,* 27: 1361-1379.
Desai, S. D., & Kouchaki, M. (2017). Moral symbols: A necklace of garlic against unethical requests. *Academy of Management Journal*, 60: 7-28
Dharmapala D, McAdams, RH (2003) The Condorcet jury theorem and the expressive function of law: a theory of informative law. *American Law and Economics Review*, 5(1): 1-31.
Ethics and Compliance Initiative (2018) *2018 Global Business Ethics Survey*. Retrieved on May 29, 2019 from http://www.ethics.org/knowledge-center/eci-recent-research/
Ethics Resource Center (2013) *National Business Ethics Survey: Workplace Ethics in Transition*. Retrieved on January 6, 2016 from http://www.ethics.org/nbes/
Fama EF (1980) Agency Problems and the Theory of the Firm. *J. of Political Economy*, 88: 288- 307.
Fehr E, Fischbacher U (2004) Social norms and human cooperation. *Trends in Cog. Sci.*, 8:185-190.
Fehr, E, Gächter, S (2002) Altruistic punishment in humans. *Nature*, 415: 137-140.
Festinger L (1954) A theory of social comparison processes. *Human Relations*, 7(2): 117-140.
Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10: 171-178.
Folger R (2001) Fairness as deonance. In Gilliland, S, Steiner, DD, Skarlicki, D (Eds). *Research in Social Issues in Management*: 3-33. New York, NY: Information Age Publishers.
Fréchette GR (2012) Session-effects in the laboratory. *Experimental Economics*, 15: 485-498.
French JR, Raven B, Cartwright D (1959) The bases of social power. *Classics of Organization Theory, 7*: 311-320.
Gellerman SW (1986) Why 'good' managers make bad ethical choices? *Harvard Bus. Rev.* 86: 85-90.
Gilovich, T., Keltner, D., & Nisbett, R. E. (2005). *Social Psychology*. New York: W. W. Norton & Company, Inc.
Gino F (2015) Understanding ordinary unethical behavior: Why people who value morality act immorally. *Current Opinion in Behavioral Sciences*, 3: 107-111.
Gino F, Ariely D (2012). The dark side of creativity: Original thinkers can be more dishonest. *J. of Per. and Soc. Psych.*, 102(3): 445-459.
Gino F, Norton MI, Weber RA (2016) Motivated Bayesians: Feeling Moral While Acting Egoistically. *The Journal of Economic Perspectives*, 30: 189-212.
Glazer E (2015) Traders' Phones Are Becoming a Surveillance Zone, *The Wall Street Journal*, (August 17), https://www.wsj.com/articles/for-traders-no-more-privacy-on-the-phone-1442168422
Gneezy U, Rustichini A (2000) A fine is a price. *Journal of Legal Studies*, 29: 1-17.
Greenwald A (1980) The totalitarian ego. Fabrication and revision of personal history, *The American Psychologist*, *25*: 603-618.
Gundlach MJ, Douglas SC, Martinko MJ (2003) The decision to blow the whistle: A social information processing framework. *Academy of Management Review*, 28: 107-123.
Hall, RC, Hall, RC (2001). False allegations: the role of the forensic psychiatrist. *Journal of Psychiatric Practice*, 7: 343-346.
Hayes AF (2013) Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (Guilford Press).
Hegarty WH, Sims HP Jr. (1978) Some determinants of unethical decision behavior: An experiment. *Journal of Applied Psychology*, 63: 451-457.

Hertwig R, Ortmann A (2001) Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24: 383-403.

Institute of Business Ethics (2015) *Ethics at Work 2015 Survey of Employees: Britain.* Retrieved from http://www.ibe.org.uk/

Jensen MC, Meckling WH (1976) Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3: 305-360.

Jones G (2012) Workplace investigations: Getting beyond 'he said, she said'. *Canadian HR Reporter,* (January 30), https://www.hrreporter.com/article/12201-workplace-investigations-getting-beyond-he-said-she-said/

Jones TM (1991) Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review*, 16: 366-395.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under uncertainty. *Econometrica*, 47: 263-291.

Kennedy JA, Schweitzer ME (2018) Building trust by tearing others down: When accusing others of unethical behavior engenders trust. *Organ. Behav. and Human Decision Processes, 149*: 111-128.

Kirchler M, Huber J, Stefan M, Sutter M (2015) Market design and moral behavior. *Man., Sci.*, 62: 2615-2625.

Knapp DE, Faley RH, Ekeberg SE, Dubois CL (1997) Determinants of target responses to sexual harassment: A conceptual framework. *Academy of Management Review*, 22: 687-729.

Lewin K (1936) *Principles of Topological Psychology* (McGraw Hill, New York).

Leung K, Tong KK (2003) Toward a normative model of justice. Gilliland SW, Steiner DD, Skarlicki DD eds. *Emerging Perspectives on Values in Organizations*, (Greenwich: Information Age Publishing), 97–120.

Louis MR (1980) Surprise and sense making: What newcomers experience in entering unfamiliar organizational settings. *Admin. Sci. Quart.*, 226-251.

Markel G, Murray H (2017) Internal investigations special committees resource. Retrieved from: https://corpgov.law.harvard.edu/2017/07/06/internal-investigations-special-committees-resource/

Martin SR (2016) Stories about values and valuable stories: A field experiment of the power of narratives to shape newcomers' actions. *Academy of Management Journal*, 59: 1707–1724.

Mažar N, Amir O, Ariely, D (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6): 633-644.

McAdams RH (2000) An attitudinal theory of expressive law. *Oregon Law Review*, 79: 339-390.

McCusker C, Carnevale PJ (1995) Framing in resource dilemmas: Loss aversion and the moderating effects of sanctions. *Organizational Behavior and Human Decision Processes*, 61: 190-201.

Miceli MP, Near, JP (1992). *Blowing the whistle: The organizational and legal implications for companies and employees* (Lexington Books).

Mitchell MS, Vogel RM, Folger R (2015) Third parties' reactions to the abusive supervision of coworkers. *Journal of Applied Psychology*, 100: 1040-1055.

Moffatt PG (2015) *Experimetrics: Econometrics for experimental economics*. Palgrave Macmillan.

Morris MW, Hong YY, Chiu CY, Liu Z (2015) Normology: Integrating insights about social norms to understand cultural dynamics. *Organ. Behv. and Human Decision Processes*, 129: 1-13.

Mueller D (2017) When is a workplace investigation necessary? *HR Acuity* (October 3), https://www.hracuity.com/blog/when-is-a-workplace-investigation-necessary

Mulder LB, Verboon P, De Cremer D (2009) Sanctions and moral judgments: The moderating effect of sanction severity and trust in authorities. *European Journal of Social Psychology*, 39: 255-269.

Murnighan JK, Wang L (2016) The social world as an experimental game. *Organ. Behav. and Human Decision Processes*, 136: 80-94.

Murphy K (2016) No one else was there: Corroboration and Bill 132, *Rubin Thomlinson LLP*, (November 24, https://rubinthomlinson.com/no-one-else-corroboration-bill-132/

Nagin, DS, & Pogarsky G (2003) An experimental investigation of deterrence: Cheating, self-serving

bias, and impulsivity. *Criminology*, 41(1): 167-194.
Nelson L (2015) Rolling Stones retracts story on alleged UVA rape. *Vox* (April 5), https://www.vox.com/2015/4/5/8347011/uva-rolling-stone-report
O'Leary-Kelly AM, Paetzold RL, Griffin RW (2000) Sexual harassment as aggressive behavior: An actor-based perspective. *Academy of Management Review*, 25: 372-388.
O'Reilly J, Aquino K (2011) A model of third parties' morally motivated responses to mistreatment in organizations. *Academy of Management Review*, 36: 526-543.
O'Reilly J, Aquino K, Skarlicki, D (2016) The lives of others: Third parties' responses to others' injustice. *Journal of Applied Psychology*, 101:171-189.
Pierce L, Snyder J (2008) Ethical spillovers in firms: Evidence from vehicle emissions testing. *Management Science*, 54(11): 1891-1903.
Podsakoff PM, Bommer WH, Podsakoff NP, MacKenzie, SB (2006). Relationships between leader reward and punishment behavior and subordinate attitudes, perceptions, and behaviors: A meta-analytic review of existing and new research. *Organ. Beh. and Human Decision Processes*, 99(2), 113-142.
Preacher KJ, Hayes AF (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40: 879-891.
Riger S (1991) Gender dilemmas in sexual harassment policies and procedures. *American Psychologist*, 46: 497-505.
Rest JR (1986) *Moral Development: Advances in Research and Theory*. Praeger Publishers.
Ryan RM, Deci EL (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1): 68-77.
Schank RC, Abelson RP (1977) *Scripts, plans, goals, and understanding: An inquiry into human knowledge*. Hillsdale NJ: Lawrence Erlbaum Associates.
Skarlicki DP, Kulik CT (2004) Third-party reactions to employee (mis) treatment: A justice perspective. *Research in Organizational Behavior*, 26, 183-229.
Skarlicki D, O'Reilly J, Kulik CT (2015) The third-party perspective of (in)justice. In Cropanzano, RS, Amborse, M (Eds), *The Oxford Handbook of Justice in the Workplace*: 235-256. Oxford.
Skinner BF (1965) *Science and Human Behavior*. Simon and Schuster.
Slater D (2010, January 25). Internal investigations: The basics. Retrieved from https://www.csoonline.com/article/2124800/investigations-forensics/investigations-forensics-internal-investigations-the-basics.html
Stephens, B (2017) When #metoo goes too far. *The New York Times* (December 20), https://www.nytimes.com/2017/12/20/opinion/metoo-damon-too-far.html
Smith-Crowe K, Tenbrunsel AE, Chan-Serafin S, Brief A., Umphress EE, Joseph J (2015). The ethics "fix": When formal systems make a difference. *Journal of Business Ethics*, *131*:791-801.
Tenbrunsel AE, Messick DM (1999) Sanctioning systems, decision frames, and cooperation. *Admin., Sci., Quart.*, 44: 684-707.
Tenbrunsel AE, Smith-Crowe K, Umphress EE (2003) Building houses on rocks: The role of the ethical infrastructure in organizations. *Social Justice Research*, 16: 285-307.
Treviño LK (1986) Ethical decision making in organizations: A person-situation interactionist model. *Academy of Management Review*, 11: 601-617.
Treviño LK, den Nieuwenboer NA, Kish-Gephart JJ (2014) (Un)ethical behavior in organizations. *Annual Review of Psychology*, 65:635-660.
Treviño LK, Weaver GR (2001) Organizational Justice and Ethics Program Follow-Through: Influences on Employees' Harmful and Helpful Behavior. *Business Ethics Quarterly*, 11: 651-671.
Tripp TM, Bies RJ, Aquino K (2007) A vigilante model of justice: Revenge, reconciliation, forgiveness, and avoidance. *Social Justice Research*, 20:10–34.
Turillo CJ, Folger R, Lavelle JJ, Umphress EE, Gee, JO (2002) Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organ., Behav., and Human Decision Processes*, 89: 839-865.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5: 297-323.

United Nations Declaration of Human Rights. Accessed February 28, 2018, http://www.un.org/en/universal-declaration-human-rights/

Valkanas G, Ipeirotis P, Provost F, Attenberg J, Chin J, Hendahewa C, Stanway A, Suryanto B, Swamy BV (2018 February). *Detecting employee misconduct and malfeasance*. Workshop paper presented at the meeting of the Association for Computing Machinery International Conference on Web Search and Data Mining, Los Angeles, USA. Retrieved from the Stanford Network Analysis Project website: http://snap.stanford.edu/mis2/files/MIS2_paper_22.pdf

Van Maanen J, Schein, EH (1979) Towards a theory of organizational socialization. Staw BM eds. *Research in Organization Behavior* 1: 209-264 (Greenwich, CT, JAI Press)

Warren DE, Smith-Crowe K. (2008). Deciding what's right: The role of external sanctions and embarrassment in shaping moral judgments in the workplace. *Research in Organizational Behavior*, 28, 81-105.

Weingast BR (1997) The political foundations of democracy and the rule of the law. *American Political Science Review*, 91: 245-263.

Yamagishi T (1986) The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51: 110-116.

Table 1. Odds Ratios from Logistic Panel Regressions of System Conditions on Behaviors (Study 1)

System Condition	Actor Stealing (1)	Recipient Reporting (2)	Observer Helping (3)
Strong System	0.17** a,b (0.11)	2.17 a (1.25)	0.73 a (0.56)
Very Strong System	0.06**a (0.03)	4.38† a,b (3.80)	5.22* a,b (4.38)
No System	0.28**b (0.13)	20.33**b (16.19)	4.84**b (2.80)
$X_2(3)$	28.28**	15.60**	9.83*
N (Clusters)	1,280 (24)	672 (24)	472 (24)

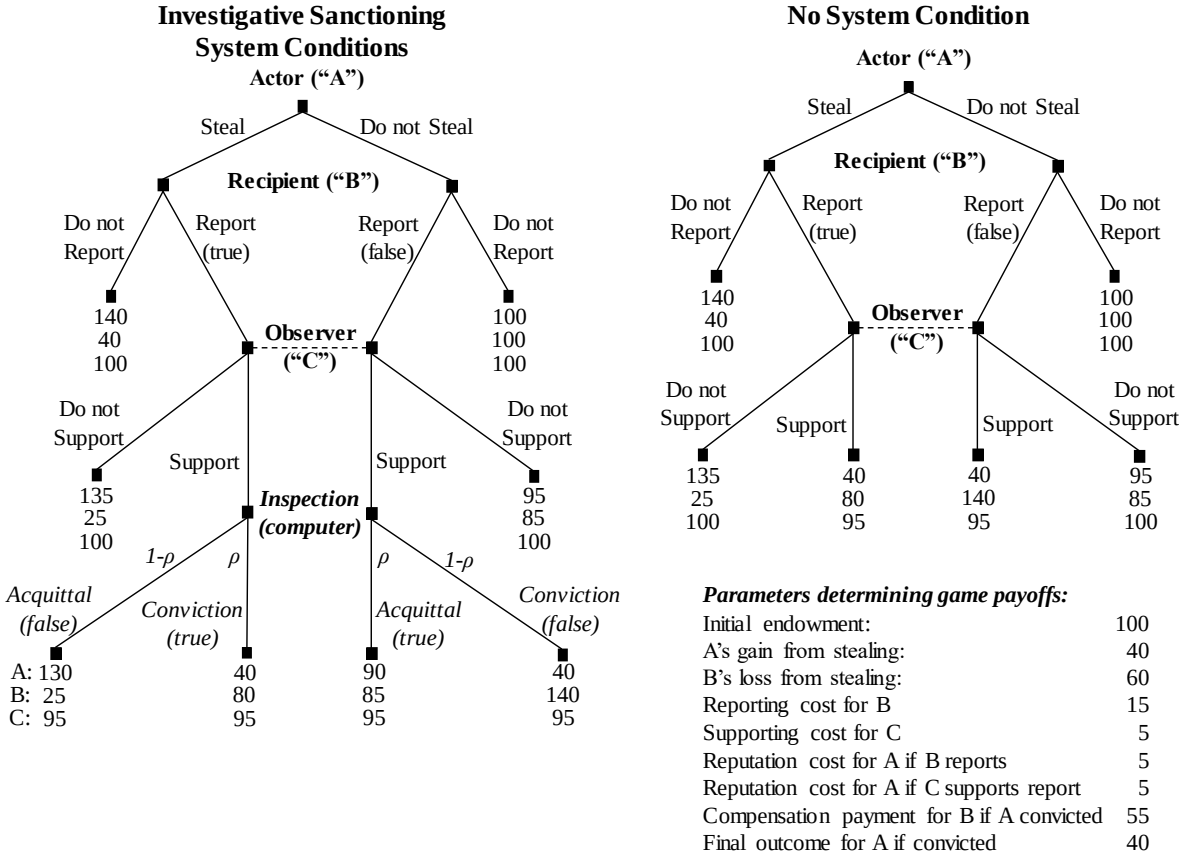
Notes: The table reports odds ratios obtained from random effects logistic panel regressions. Robust standard errors are in parentheses, clustered by 24 sessions. The weak system condition is the omitted base category. An odds ratio below 1 means that the likelihood of engaging in the respective behavior is smaller than in the weak system condition, an odds ratio above 1 means that it is greater. Column (2) only includes cases where the actor stole points and column (3) only includes cases where the recipient made a report, and the observer received a (possibly false) computer message that the actor had stolen points. Odds ratios within the same column that do not share a subscript are significantly different from each other ($p<.05$). Note that for contrasts between the strong, the very strong, and the no system condition, the *p*-values reported in the text stem from post-estimation Wald tests based on the regressions reported in this table. The superscript significance markers indicate significant differences compared to the omitted weak system condition (i.e., an odds ratio significantly different than 1): †$p < .10$, *$p < .05$, **$p < .01$

Table 2. Odds Ratios from Logistic Panel Regressions of System Condition on Behaviors (Study 2)

System Condition	Actor Stealing (1)	Recipient Reporting (2)	Observer Helping (3)
Very Strong System	0.07** (0.04)	1.98 (1.14)	7.50* (5.98)
$X_2(3)$	18.18**	1.40	6.39*
N (Clusters)	1,200 (24)	634 (24)	403 (24)

Notes: The table reports odds ratios obtained from logistic panel regressions. Robust standard errors are in parentheses, clustered by 24 sessions. The weak system condition is the omitted base category. An odds ratio below 1 means that the likelihood of engaging in the respective behavior is smaller in the very strong than in the weak system condition, and an odds ratio above 1 means that it is greater. Column (2) only includes cases where the actor stole points and column (3) only includes cases where the recipient made a report, and the observer received a (possibly false) computer message that the actor had stolen points. The superscript significance markers indicate significant differences compared to the omitted weak system condition (i.e., an odds ratio significantly different than 1): †$p < .10$, *$p < .05$, **$p < .01$

Figure 1. Decision Trees of Experimental Game (Studies 1 and 2)

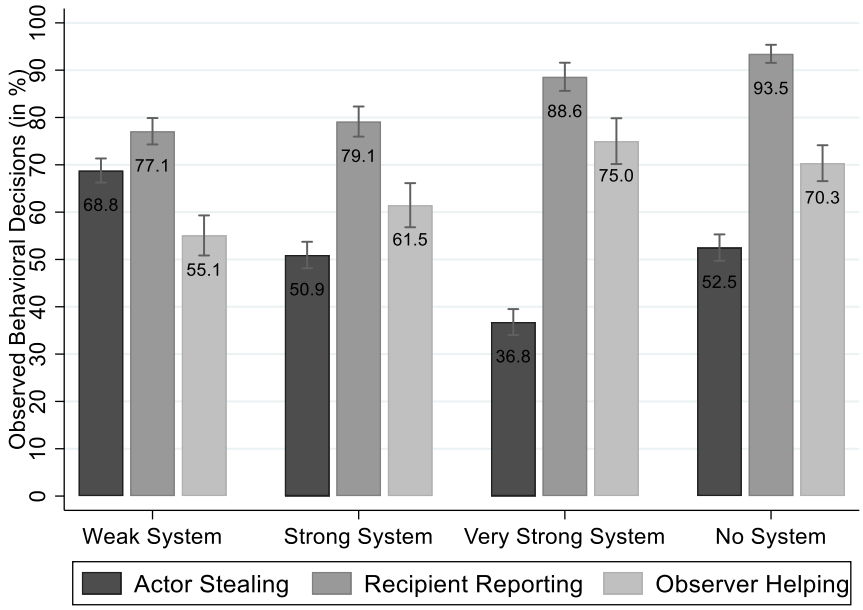


Parameters determining game payoffs:

Initial endowment:	100
A's gain from stealing:	40
B's loss from stealing:	60
Reporting cost for B	15
Supporting cost for C	5
Reputation cost for A if B reports	5
Reputation cost for A if C supports report	5
Compensation payment for B if A convicted	55
Final outcome for A if convicted	40

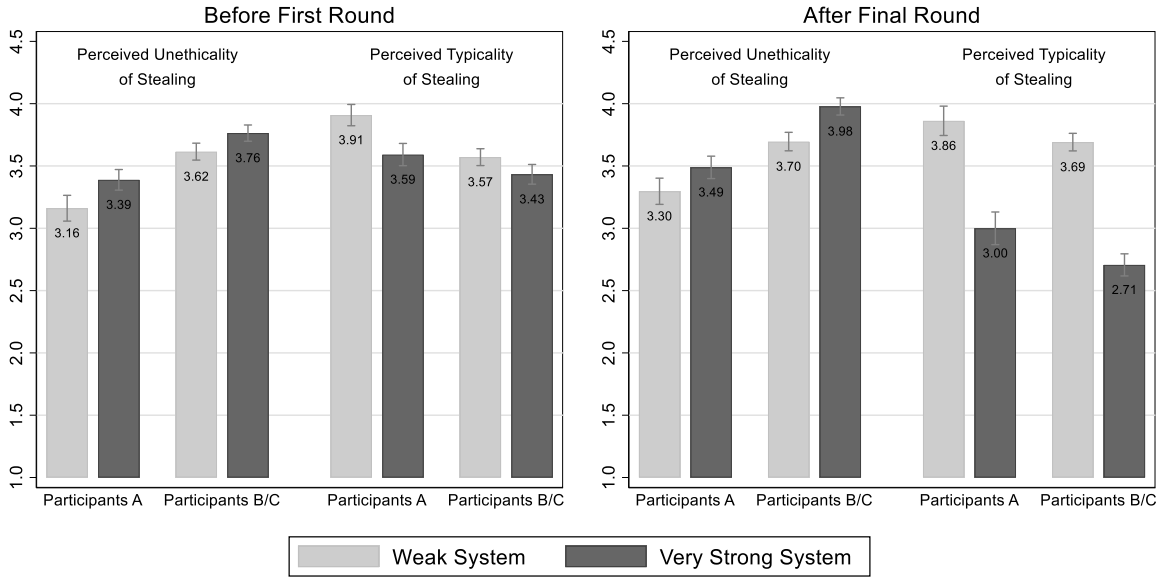
Notes: Final payoffs for the actor ("A", top row), the recipient ("B", middle row), and the observer ("C", bottom row) are displayed at all possible end-nodes of the game. The exchange rate of points to CHF was 9 to 1 in Study 1 and 10 to 1 in Study 2. The outcomes of two randomly chosen rounds (one from the first five rounds, one from the second five rounds) were paid out in cash at the end of the experiment. The dashed line depicts the information-set of C. When making a helping decision, C only knows that B has made a report, but not whether A had decided to steal or not. Before making the decision, C receives an imperfect computer message (with an 80% chance of being correct) indicating whether A had stolen. The computer message is not depicted above to simplify the illustration. Parameter ρ refers to the probability that the inspection in the surveillance and sanctioning system conditions correctly uncovers the truth (i.e., whether A stole or not). ρ=55% in the weak, ρ=85% in the strong, and ρ=100% in the very strong system condition. In all experimental conditions the backward-induction equilibrium with selfishly money-maximizing and rational players is that A steals, and B does not report, because C never supports.

Figure 2. Behavioral Decisions across Experimental Conditions (Study 1)



Notes: The figure shows average rates of stealing (N_{Weak}=330, N_{Strong}=320, $N_{VeryStrong}$=310, N_{No}=320 decisions), reporting (i.e., recipients reporting in cases where the actors had actually stolen points; N_{No}=168, N_{Weak}=227, N_{Strong}=163, $N_{VeryStrong}$=114 decisions), and helping (i.e., observers supporting a recipient's report when they had received a computer message indicating that the actor had stolen; N_{No}=181, N_{Weak}=195, N_{Strong}=174, $N_{VeryStrong}$=132 decisions) in Study 1. Error bars represent plus/minus one binomial standard error (unclustered).

Figure 3. Moral Judgment and Descriptive Norm Perceptions of Stealing (Study 2)



Notes: The figure shows the average moral judgment and descriptive norm perceptions of stealing as measured before the first (left panel) and after the final round (right panel) of the experiment in Study 2. N_{Weak}=180 and $N_{VeryStrong}$=180 participants (with 60 participants A and 120 participants B/C in both conditions). Error bars represent plus/minus one standard error of the mean (unclustered).