Available online at
http://www.anpad.org.br/bar

BAR, Rio de Janeiro, RJ, Brazil v. 15, n. 4,
art. 3, e180016, 2018
http://dx.doi.org/10.1590/1807-7692bar2018180016

Explaining transgressions with moral disengagement strategies and their effects on trust repair

Tatiana Iwai[1]
João Vinícius de França Carvalho[2]
Victor Marson Lalli[1]

Insper Instituto de Ensino e Pesquisa, São Paulo, SP, Brazil[1]
Universidade de São Paulo, Faculdade de Economia, Administração e Contabilidade, São Paulo, SP, Brazil[2]

Received 1 February 2018. This paper has been with the authors for two revisions. Accepted 10 December 2018. First published online 19 December 2018.
Editor's note. Filipe João Bera de Azevedo Sobral served as Associate editor for this article.

Abstract

When providing explanations for a transgression, the offender may use verbal statements based on moral disengagement strategies to mitigate the negative consequences of a trust violation. That is, the offender may try to reframe unethical acts to appear less harmful, or displace responsibility for the wrongdoing, or distort the consequences of his or her actions in order to address the repair of a damage of a trust violation. Based on this, we examined the effects of these explanations based on different moral disengagement strategies on trust repair. The results of a scenario-based experiment show that both the moral justification and the displacement of responsibility strategies elicited higher trusting intentions compared to the distortion of consequences strategy. This effect was mediated by trusting beliefs toward the offender. These findings suggest that reframing the unethical conduct as targeting a greater good, as well as obscuring personal agency for the detrimental conduct, may be more effective to repair trust than misrepresenting the consequences of the immoral acts.

Key words: moral disengagement strategies; explanations; trusting intentions; trusting beliefs.

Introduction

Although trust is associated with a broad array of positive outcomes for individuals, groups, and organizations (Dirks & Ferrin, 2001), it is considered fragile and easily broken. For this reason, many scholars have been paying attention to understand how trust can be repaired, once broken. Among the reparative efforts to restore trust, studies point out verbal strategies as a possible approach to mitigate negative responses in the aftermath of a trust violation (Lewicki & Brinsfield, 2017).

More than acknowledging that a violation has occurred, a verbal statement aims to provide an explanation for the violation, revealing the underlying reason or the cause of the negative event. For instance, imagine the following situation: you are told someone had made inappropriate use of corporate resources. Would it make any difference if that person claimed to have done to help someone? Or if they asserted that an authority figure told them to do this? Or if they allege there was no big deal because no one got hurt? Would your perception about that person change based on these different explanations? Would your willingness to make yourself vulnerable to that person be greater or lower according to those explanations?

The language and content of those verbal accounts are consistent with what Bandura's social cognitive theory (Bandura, 1986, 1999; Bandura, Barbaranelli, Caprara, & Pastorelli, 1996) calls moral disengagement mechanisms. They are cognitive mechanisms individuals employ to persuade themselves that an unethical behavior is morally acceptable and, as a consequence, allow individuals to act unethically and still feel moral. Although these mechanisms are usually depicted as strategies used by individuals to justify their own misdeeds, previous studies have been arguing that these cognitive mechanisms can also be considered as possible statements to explain misconduct for others, either victims or observers (Dang, Umphress, & Mitchell, 2017; Reynolds, Dang, Yam, & Leavitt, 2014).

Moral disengagement mechanisms work by decoupling one's internal moral standards from one's acts. That is, they selectively disable our moral self-regulatory processes that restrict us from acting inhumanly. It is worth noting that Bandura (1986, 1996) clusters these mechanisms in distinctive groups, according to the different points in the self-regulatory system at which internal moral control can be disengaged from detrimental conduct. Moral self-censure can be disengaged from immoral behavior by reconstructing the action into a benign one, minimizing personal agency for the conduct, disregarding or misrepresenting the consequences of action, or devaluing the recipients of detrimental acts.

In this research, we focus on the mechanisms from the first three groups of disengagement practices: (a) moral justification, (b) displacement of responsibility, and (c) distortion of consequences. That is, an offender can justify his or her misbehavior by (a) appealing to higher order goals, (b) deflecting responsibility for the wrongdoing to someone else, or (c) minimizing the injurious consequences of the detrimental action. We aim to examine how explanations based on these different moral disengagement strategies affect trust repair. We designed an experiment in which we manipulated different explanations for a previous unethical behavior and compared individuals' responses to these verbal statements.

Our contributions are threefold. First, we advance the moral disengagement literature. Previous studies have usually described moral disengagement as an internal cognitive process leading to unethical conduct (Detert, Treviño, & Sweitzer, 2008; Martin, Kish-Gephart, & Detert, 2014; Moore, Detert, Treviño, Baker, & Mayer, 2012). We expand this understanding by considering moral disengagement strategies as possible explanations to reframe unethical behavior as less reprehensible not just for ourselves, but in the eyes of others. By doing that, we expand the application of moral disengagement to a broader social context (Dang et al., 2017). Second, we also contribute to the verbal allegations and impression management literature. By using moral disengagement strategies as possible rationales that individuals employ to explain trust violations, we shifted attention from verbal allegations extensively studied as denials and apologies (Lewicki & Brinsfield, 2017) to other forms of explanations that still remain underexplored in the literature (Kim & Harmon, 2014; Shaw, Wild, & Colquitt, 2003). Finally, we add to the literature on trust repair by showing what types of explanations using ethically oriented

language can be more effective to repair trust after an integrity-based violation (Kim, Dirks, Cooper, & Ferrin, 2006; Kim, Ferrin, Cooper, & Dirks, 2004).

In the sections that follow, we first set out the foundations of our arguments by reviewing verbal allegation literature, as well as Bandura's moral disengagement theory, and then develop our hypotheses. Next, we detail our research methodology and present the results of our analysis. We conclude with a discussion of our findings, as well as limitations of the study and future research possibilities.

Theory and Hypotheses

Verbal allegations and trust repair

Although there are many definitions for trust, we can define it "as a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another" (Rousseau, Sitkin, Burt, & Camerer, 1998, p. 395). In this definition, trust includes cognitive, affective and behavioral intention elements, which may all be damaged or depleted as a result of a trust violation. Most importantly, according to this definition, vulnerability is an essential element of trust. It means that trust comprises a behavioral intent component related to risk, since making oneself vulnerable to the actions of another is willingness to take risks at the hands of others (Mayer, Davis, & Schoorman, 1995). As Lewicki, Tomlinson, and Gillespie (2006, p. 998) stated, "to trust behaviorally involves undertaking a course of risky action based on the confident expectations (cognitive basis) and feelings (emotional basis) that the other will honor trust".

Therefore, in the aftermath of a trust violation, willingness to make oneself vulnerable to another party will be damaged by the prior negative outcome of risk taking. However, negative judgments following a transgression are not necessarily final. Prior research has shown how those expectations and dispositions to take risks at the hands of others can be modified or improved by explanations of the causes of that incident (Kim et al., 2006, 2004; Kim & Harmon, 2014; Tomlinson, Dineen, & Lewicki, 2004; Weiner, 1985). Specifically, they can shape the other's perception about a transgression by revising damaged attributions in the aftermath of a trust violation (Tomlinson & Mayer, 2009). Verbal allegations include denials, apologies, explanations and every other verbal statement used by the transgressor to address the violation and restore the relationship (Lewicki & Brinsfield, 2017).

Among them, denials and apologies have attracted more attention from scholars, who have been particularly interested in examining which one is more effective to repair broken trust (Kim et al., 2004, 2006; Lewicki & Brinsfield, 2017). Some scholars have argued that apologies, by acknowledging responsibility and regret for the wrongdoing, convey a sincere intention to behave differently in the future, which should reduce victim's perceptions of future vulnerability and, as a consequence, may repair trust more effectively (Kim, Cooper, Dirks, & Ferrin, 2013; Tomlinson et al., 2004). Other scholars, though, have argued that apologies may backfire, due to their acknowledgement of guilt, which might exacerbate the detrimental consequences of an accusation (Kim et al., 2004; Schlenker, 1980). A denial, conversely, states that the alleged violation is untrue and, in doing so, exempts oneself of any responsibility for the harm inflicted. By rejecting any blame for the act, the accused party may eliminate any need to repair his or her trustworthiness, if those claims were deemed true (Kim, Dirks, & Cooper, 2009).

Extant evidence suggests that a denial can be a particularly effective response, according to the nature of the trust violation. Whereas apologies can be more effective when violations concern matters of competence, they are inferior responses in comparison to denials when violations concern matters of integrity (Ferrin, Kim, Cooper, & Dirks, 2007; Kim et al., 2004, 2006). Such findings can be explained by the differences in how we weigh positive and negative information regarding competence and integrity matters (Snyder & Stukas, 1999). Although individuals, when considering matters of competence, tend to weigh positive information more heavily than negative information, the opposite

tendency occurs to matters of integrity (Kim, Diekmann, & Tenbrunsel, 2003). For instance, for competence matters, a single episode of outstanding performance is enough as a reliable signal of competence, since those who are incompetent would not be able to perform at that high level.

However, for integrity matters, a single dishonest behavior provides a strong signal of lack of integrity, since an honest person will refrain from behaving dishonestly in any situation. As a result, the costs of using an apology for an integrity-based violation (due to the confirmation of guilty) may outweigh its benefits (due to potential redemption), because it takes only one single episode of lack of integrity to create a permanent perception of defective character in the transgressor. As Ferrin, Kim, Cooper and Dirks (2007, p. 895) assert, "this is because people tend to believe that a lack of integrity would only be exhibited by those who do not possess it, and this belief, once established, is difficult to disconfirm".

Despite their efficacy, in many situations, denials can be problematic, simply because the mistrusted party could have actually committed the transgression and, in such cases, not only would it be unethical to deny culpability, but it also might be counterproductive if evidence of guilt come up afterwards (Kim et al., 2004, 2009; Kim & Harmon, 2014). In those circumstances, scholars pointed out the importance of giving explanations for the detrimental behavior in order to provide relevant contextual information to clarify the underlying causes of the violation (Shaw et al., 2003). However, those explanations will only be able to mitigate the negative consequences of a transgression, if perceivers judge the explanations to be adequate (Shapiro, 1991). Therefore, those explanations should employ content and language that make them not only more credible and reasonable in the eye of the beholder, but also provide extenuating circumstances for the transgression in order to alleviate the adverse social effects of a trust violation.

Moral disengagement theory

One particularly important theoretical framework to shed light on how individuals explain misbehavior is moral disengagement theory (Bandura, 1986, 1999; Bandura et al., 1996). It is an extension of Bandura's social cognitive theory, which states that people develop personal standards of right and wrong that serve as guides and deterrents of moral conduct. In this self-regulatory system, behaviors are reinforced or deterred according to how they are judged in relation to our internalized moral standards. Actions in line with internal standards will bring individuals a sense of satisfaction and self-worth, while actions that violate moral standards will induce feelings of self-condemnation such as guilt, remorse and shame. Therefore, people usually behave in accordance with their moral standards, because they anticipate positive and negative self-reactions they will face as a result of the conduct they choose.

Moral standards do not operate as invariant internal regulators of conduct, though. They have to be activated to deter immoral behavior. However, there are many psychological maneuvers to disengage self-sanctions from detrimental conduct. According to Bandura (1996, 1999), self-regulatory processes, which normally inhibit us from making unethical decision, can be selectively activated or deactivated by what he calls moral disengagement mechanisms. They refer to cognitive mechanisms that decouple the cognitive links between unethical behavior and the moral self-censure that should prevent it. By disengaging from self-sanctions, individuals are freed from the anticipatory negative emotions they experience when behaving in ways that violate internal moral standards.

Bandura (1986, 1996, 1999) identifies eight cognitive mechanisms to describe an immoral act in such a way that makes it appear less wrong and, as a consequence, allows the subject to behave unethically and still feel moral. They can be clustered into four sets of moral disengagement strategies that operate at different points in the self-regulatory system at which moral self-sanctions can be disengaged from immoral conduct (Bandura, 2011; Bandura et al., 1996).

In the first group, the disengagement has a behavioral locus and concentrates on cognitively reframing harmful conduct as more morally acceptable. It consists of three mechanisms. Moral justification reconstrues reprehensible behavior as serving the greater good. For instance, studies show

that many policy officers, when forced to choose between lying under oath (perjury) and testifying against their colleagues, prefer the first option and justify the act as loyalty to their peers (Anand, Ashforth, & Joshi, 2004). With euphemistic language, individuals use sanitized language to describe a conduct in a way that appears less harmful or wrong. Killing civilians in war may be rephrased as collateral damage, for example (Bandura, 1999). Advantageous comparison involves comparing a wrongdoing with a worse behavior, making the first behavior appears more acceptable, as in the case of a student comparing the act of cheating on an exam with submitting academic work done by someone else.

The second set of disengagement practices has an agency locus and operates by obscuring personal agency for the harmful act in two ways. With displacement of responsibility, individuals deflect responsibility for a wrongdoing by attributing it to authority figures that dictate that they should engage in unethical behavior (e.g., my boss told me to do it). Diffusion of responsibility is another way of mitigating agentive role in the harm inflicted. It works by spreading responsibility among group members engaging in the same adverse conduct. When everyone is involved in the misconduct, no one feels personally liable for it (e.g., everyone is doing it too).

Another way of weakening moral self-sanctions has an outcome locus and works by distorting or minimizing the injurious consequences of misbehavior. With distortion of consequences, self-censure is deactivated due to misrepresentation of the outcomes of one's misbehavior. They are distorted or ignored in a way that misconduct seems harmless. This is commonly employed in situations when the victim is powerful or wealthy. For instance, when an employee commits a minor theft from his or her large and profitable employer (Anand et al., 2004).

Finally, two moral disengagement mechanisms focus on the victim and operate by devaluating or blaming the victim of the maltreatment. They aim to reduce identification with the target of the harmful acts. In attribution of blame, victims are blamed to have brought the harm on themselves. For example, claiming that "if people have their privacy violated, it's probably because they have not taken adequate precautions to protect it" (Moore et al., 2012, p. 48). Lastly, dehumanization involves devaluing the victim as not deserving human basic consideration. By divesting the victims of basic human qualities, our identification with the victim lessens and our moral self-sanction is less likely to be activated. Bandura (1999) notes how nations, in war times, try to degrade the enemies to a subhuman status to make it easier to kill them.

Although Bandura primarily articulated moral disengagement as a process, many scholars have usually approached it as a dispositional tendency (Detert et al., 2008; Moore et al., 2012; Moore, 2015). For that reason, empirical works typically measure it as "an individual difference in the way that people cognitively process decisions and behavior with ethical import that allows those inclined to morally disengage to behave unethically without feeling distress" (Moore et al., 2012, p. 2). However, more recently, there is a growing interest in considering how situational characteristics may trigger moral disengagement mechanisms (Barsky, 2011; Gino & Galinsky, 2012; Kish-Gephart, Detert, Treviño, Baker, & Martin, 2014; Shu, Gino, & Bazerman, 2011). These studies provide support for the argument that moral disengagement is a motivated cognitive process. That is, when it is in the individual's interest to disengage from self-moral standards, they are more likely to do so (Martin et al., 2014; Moore, 2015; Tsang, 2002). As Tenbrunsel and Messick (2004) argue, an ethical decision usually involves a trade-off between one's self-interest and moral standards to follow. However, individuals can solve this conflict of motivations and their need to see the self as moral by using rationalizations to reframe their immoral behaviors as, in fact, moral. Such psychological maneuvers allow individuals to behave in a self-interested manner and still convince themselves that their moral principles were upheld (Tsang, 2002).

Consistent with this reasoning, we can theorize that in the same way that moral disengagement is strategically used to clear conscience about one's own misdeeds, it also can be employed as an effective tactic to manage other's impressions about one's transgressions. Although currently in limited number, some previous studies had already departed from depicting moral disengagement mechanisms only as strategies to justify their own misbehavior and started to consider them as possible statements used to

explain unethical conduct for others, either victims or observers (Dang et al., 2017; Reynolds et al., 2014).

Such understanding is in line with theories of impression management (Bolino, Long, & Turnley, 2016) and self-presentation (Baumeister, 1982), which argue that not only do individuals care about their self-image, but they are also concerned with managing the image others have of them. Therefore, following a transgression, individuals could use moral disengagement to see themselves as moral, as well as to present themselves as such to a target audience.

Therefore, we also consider moral disengagement (MD) mechanisms as possible explanations to justify one's transgressions to others and mitigate the adverse effects of a trust violation. In the current study, three MD strategies were chosen: moral justification, displacement of responsibility and distortion of consequences. We focus on them, because each one is grounded on different loci at which moral sanctions can be disengaged from detrimental conduct: moral justification has a behavioral locus, displacement of responsibility focuses on obscuring the agent and, finally, distortion of consequences operates based on an outcome locus. Therefore, we chose one strategy of each set of disengagement practices and excluded those with a victim locus, because they seem less relevant to an organizational setting. Moreover, specifically about the choice of moral justification and displacement of responsibility, we can also argue that not only have they both been widely researched (Barsky, 2011), but they also provide an interesting opportunity to compare explanations based on different levels of intention. While displacement of responsibility attempts to shift or minimize personal accountability for the conduct, moral justification explicitly assumes more responsibility for the act, but tries to reframe it in a positive perspective. Based on these choices, we compare the effects of these three types of moral disengagement strategies on trust repair.

Repairing trust with explanations based on moral disengagement strategies

In their bilateral model of trust repair, Kim, Dirks and Cooper (2009) contend that trust repair is the result of a negotiated process between trustors and trustees to resolve discrepant beliefs about trustee's trustworthiness. Following a trust violation, while the trustee wants to be considered trustworthy, the trustor may resist believing that greater trust in the trustee is deserved. According to the model, such discrepancies can be solved on multiple levels, which represent a sequence of questions about if the accused party is trustworthy: (a) is the accused party guilty or not of committing the transgression?; (b) if the accused party is guilty, should the underlying causes be attributed to the person or to the situation?; (c) if they are at least partially attributed to the person, is that personal cause stable or not?

According to the model, each successive question represents a less comprehensive level of trust repair. For instance, the broadest level of trust repair possible is to be considered completely innocent, which is what a denial intends to aim at. However, when it is not feasible to be completely exonerated from the accusations, one may target efforts to get a partial trust repair, even if it is narrower in scope. The accused party may then attempt to deflect part of the blame to the situation and convince the observer that the act was induced by some situational factor. Finally, when it is not possible to mitigate even part of the culpability, the accused party may then try to repair trust by claiming the detrimental conduct will not happen again in the future and, in doing so, reduce concerns about future violations. An apology, by acknowledging responsibility and expressing regret for a transgression, is a possible approach to succeed in this narrower-scope trust repair (Ferrin et al., 2007; Kim et al., 2004, 2006).

Interestingly, all these considerations can be understood as part of an attempt to make sense of the negative event. That is, such questioning helps the process of making attributions about the causes of a negative event. In this sense, attribution theory (Heider, 1958; Kelley, 1967; Weiner, 1986) can be particularly useful to explain the different effects explanations based on moral disengagement strategies may have on trust repair.

According to an attributional perspective, trustees will try to shape trustor's attribution about whether they can be blamed or not for the transgression and if this event can predict trustee's actions in

future interactions as well (Dirks, Lewicki, & Zaheer, 2009; Elangovan, Auer-Rizzi, & Szabo, 2007). For that, according to Weiner's causal attribution theory (1986), trustors evaluate a negative event by making three important attributions. First, they make an attribution of causality, in which they assess who or what caused the event. As a result, the event may be attributed to the trustee (internal) or external to him or her. Second, there is an attribution of controllability, in which trustors evaluate the level of control trustees had over the negative outcome. Such attribution assesses how much trustees can be considered accountable or not for the negative event (Shaver, 1985). Finally, attribution of stability refers to the extent the cause can be considered stable or not and, as such, it can predict what can be expected from the trustee in future interactions. Jointly, these attributions will be used to draw inferences about trustee's traits and intentions. Consequently, the results of these attributions may support or disconfirm the negative impression caused by a trust violation (Dirks et al., 2009; Tomlinson & Mayer, 2009).

Building on these considerations, when comparing explanations based on moral justification, displacement of responsibility and distortion of consequences, we argue that the latter is an inferior response in comparison to the former ones. Displacement of responsibility, by shifting accountability for a wrongdoing to someone else (e.g., management orders, peer pressure), clearly aims to minimize one's personal responsibility for the transgression. By claiming that behavior is due to circumstantial factors, displacement of responsibility is able to reduce internal attributions, controllability attributions, and stability attributions. Attenuating culpability has been indeed found to prove beneficial in many aspects by reducing anger (Weiner, Amirkhan, Folkes, & Verette, 1987), as wells as willingness to punish transgressor (Shaw et al., 2003; Wood & Mitchell, 1981), for instance.

As to moral justification, it works by accepting responsibility for the wrongdoing, but portraying it in the service of worthy or moral purposes. It tries to convince the trustor that the harmful act was actually appropriate due to a social norm or value that the trustee felt morally obliged to comply with. The harmful act is then justified as serving a social or moral imperative. Interestingly, such line of justification would strengthen internal, controllable and stable attributions to misbehavior. As a result, it would damage trust. However, as moral justification revisits the action through positive lens, instead of hurting trust, this explanation actually helps improve it. Such a strategy has also been found to be effective in reducing sanctions, retaliations and gaining social approval (Shapiro, 1991; Shaw et al., 2003).

In contrast, an explanation based on distortion of consequences tries to convince the observer the detrimental act is harmless by disregarding or distorting its effects. Therefore, contrary to displacement of responsibility, such explanation neither denies nor displaces culpability for the situation by alleging that external forces played a role in the situation. As a result, it fails to disconfirm guilt, which may lead to internal, controllable, and stable attributions for the transgression. Moreover, in comparison to moral justification, it is also a suboptimal response, because diminishing the negative consequences of the transgression does not offer legitimate and high moral values reasons for the wrongdoing, as the other strategy does. As a result, one might expect that such explanations may be unable to succeed even in the narrower-scope trust repair attempts. Thus, we hypothesize the following:

> **Hypothesis 1a:** Explanations based on moral justification or based on displacement of responsibility lead to higher trusting intentions than explanations based on distortion of consequences.

In spite of the benefits yielded by a moral justification strategy, because it assumes more responsibility and intention for a wrongdoing compared to a displacement of responsibility strategy, it is reasonable to expect that it may hinder trust more heavily, especially in integrity-based trust violations, where assuming guilt may be particularly more detrimental, as previously mentioned. Along similar lines, in a meta-analysis of studies focused on excuses (less responsibility) and justifications (more responsibility), Shaw, Wild and Colquitt (2003) found some support that excuses should have more beneficial effects than justifications.

However, scholars have been recently arguing that there is a broader approval of unethical behavior that helps others (Gino, Ayal, & Ariely, 2013; Gino & Pierce, 2009). For instance, scholars argue that deception harms trust, because it involves intentionally misleading others (Croson, Boles, & Murnighan, 2003; Schweitzer, Hershey, & Bradlow, 2006). However, recent studies have shown that, depending on the intention of the deception, lies can actually increase trust. When the deception aims to benefit the target, what scholars called prosocial lies (Levine & Schweitzer, 2014) or white lies (Erat & Gneezy, 2012), then there are positive consequences in behavioral and attitudinal measures of trust (Levine & Schweitzer, 2015).

According to Bandura, Barbaranelli, Caprara and Pastorelli (1996), moral justification is considered the most powerful mechanism for disengagement of self-sanctions, since reconstructing harmful acts into good ones not only weakens self-censure, but also transforms the misbehavior into a source of positive self-valuation. Thus, although an explanation based on moral justification admits responsibility for the wrongdoing, by framing it in an altruistic perspective and in accordance with some type of superordinate goal, it may not only mitigate the negative aspects of the unethical act, but it may also recast the act as a signal that the other holds high moral standards (Gino et al., 2013; Levine & Schweitzer, 2014). Therefore, by using moral justification, trustees induce trustors to believe that their actions are aligned with their high moral values. Such perception conveys a message of predictability of behavior in future interactions, which is a fundamental facet of trust (Peus, Wesche, Streicher, Braun, & Frey, 2012). By contrast, displacement of responsibility is not able to reverse the negative action into a positive one, as it only tries to deflect part of the responsibility for the wrongdoing to external causes. Based on this, we hypothesize the following:

> **Hypothesis 1b:** Explanations based on moral justification lead to higher trusting intentions than explanations based on displacement of responsibility.

As stated previously, trust is a psychological state that entails willingness to be vulnerable to another in risk situations due to positive expectations towards the other's behaviors and intentions. By this definition, and consistent with McKnight, Cummings, and Chervany (1998), we consider that trust encompasses not only trusting intentions, but also trusting beliefs and that the latter influence the former. That is, trust entails not only a behavioral intention component (willingness to make oneself vulnerable to another), but a cognitive one as well (beliefs about one's trust-relevant qualities, such as beliefs about other's competence or integrity that may lead to trusting intentions).

As such, in line with the literature that links beliefs to intentions (Ajzen, 2012; Fishbein & Ajzen, 1975), those trusting beliefs can be considered the foundations from which trusting intentions are grounded, as this cognitive basis of trust allows for the reduction of uncertainty regarding another's conduct. As theorized by some scholars (Kim et al., 2004, 2009; Mayer et al., 1995; McKnight, Cummings, & Chervany, 1998; Tomlinson & Mayer, 2009), the beliefs about an individual should affect one's intentions to act in a particular way toward that other.

Therefore, particularly important in any trust repair process are the perceptions of some characteristics of the trustee. They form the basis for the expectations of how one will behave in the future (Lewicki, Tomlinson, & Gillespie, 2006). Specifically, three characteristics of that trustee may allow one to assess his or her level of trustworthiness and create expectations about his or her future behavior (Mayer et al., 1995): ability (perceptions of the other's competence to perform up to expectations), benevolence (perceptions of the other's goodwill) and integrity (perceptions of the other's moral and ethical principles and standards).

Each of them contributes to explain how much one trusts another. However, it is worth noting that although each dimension is important, they may vary independently of the others (Mayer & Davis, 1999; Mayer et al., 1995). It means that it is important to identify which dimension of trustworthiness was damaged and needs to be restored. In this sense, when a trust violation occurs, the efforts to repair trust have to be concentrated on improving the trustworthiness dimension that declined (Kim et al., 2009; Tomlinson & Mayer, 2009). For instance, if one's low integrity is perceived as the cause of a

negative event, trying to appear more competent to compensate for that negative impression of character will be of little help to repair trust.

Accordingly, in cases of moral transgressions, which are an integrity-based trust violation (Lewicki & Brinsfield, 2017), explanations based on moral disengagement strategies can be particularly useful to repair trust. That is because they will specifically target the improvement of the trustworthiness dimension related to one's integrity. Therefore, we argue that, when an integrity-based trust violation occurs, the explanations based on moral disengagement strategies will aim to reestablish damaged perceived integrity (trusting beliefs) following the incident. It may, in turn, lead to higher willingness to make oneself vulnerable to another (trusting intentions).

However, as mentioned previously, although all moral disengagement strategies attempt to minimize or remove the unethical content of unethical behavior, each of them achieves this by employing distinct justifications for the wrongdoing. As such, they may influence perceived integrity differently, which, in turn, may lead to different levels of trusting intentions. Therefore, the indirect effect of the explanations based on moral disengagement strategies should influence trusting intentions (willingness to make oneself vulnerable to another) through trusting beliefs (perceived integrity).

Aligned with our previous hypotheses, among the three moral disengagement strategies (moral justification, displacement of responsibility, and distortion of consequences), the first will lead to higher perceived integrity damaged by the transgression than the latter ones. That is because explanations based on moral justification reframe detrimental behaviors into altruistic ones. As a result, they reverse the negative content of the action into a positive and highly value-based one, which helps to improve one's perceived integrity. As for the others, they fail to communicate that, because they either attenuate responsibility for the unethical behavior by shifting it to someone else, or justify misdeeds by diminishing the consequences of one's misbehaviors. In the same vein, comparing distortion of consequences with displacement of responsibility, the latter justification is more likely to lead to one's higher perceived integrity. This happens because at least displacement of responsibility attempts to deflect personal accountability for the negative event to situational factors. That shift may help to preserve perceived integrity by demonstrating that the cause of the negative outcome was less controllable and more unstable (Tomlinson & Mayer, 2009; Weiner et al., 1987). In contrast, just disregarding the consequences of the misbehavior fails to show that the misbehavior was out of his or her volitional control, as well as signals that this negative behavior is more stable and that one can expect the same conduct in future interactions.

Thus, even through different means, all these explanations aim to improve perceived integrity, which, in turn, may influence trusting intentions (Mayer et al., 1995; McKnight et al., 1998). Therefore, one might expect that the effects of type of explanation on trusting intentions will be mediate by perceived integrity. Thus, we hypothesize the following:

> **Hypothesis 2:** Perceived integrity mediates the relationship between type of explanation and trusting intentions.

Method

We test these hypotheses in a scenario-based experiment, in which we manipulated the explanations provided by the transgressor for a wrongdoing and examined perceived integrity resulting from each explanation, as well as trusting intentions toward the transgressor.

Sample and procedure

One hundred and thirty-nine senior undergraduate students and recent graduates from accounting and actuarial sciences courses participated in this study. The average age was 23.66 years, with 35.65

months of work experience, and 46.00% of them were female. Participants were randomly assigned to the three study conditions.

Based on a setting proposed and applied previously by Kim, Ferrin, Cooper and Dirks (2004), the material used in this study asked participants to assume the role of a manager in charge of hiring, and subsequently managing, an accountant. Participants read a vignette, in which they were told that, in order to expedite the hiring process, initial interviews had already been conducted by the HR department. Based on these interviews, they prepared a short report highlighting the most important attributes of each candidate, so that the manager (the participant's role) can quickly assess them. One of the candidates, Mario, met all the required qualifications for the job regarding years of experience in the position, academic background and previous job responsibilities. Moreover, he had been positively evaluated in psychological and attitudinal aspects, as well as in terms of fitting the organizational culture. However, the HR report also informed that, after contacting his previous employer, he mentioned that Mario had been involved in an incident of adulteration of the company's financial results. The recruiter then mentioned the incident to the candidate in order to give him the opportunity to provide an explanation for what had happened in that occasion. Participants then read the candidate's explanations for the incident so that they could provide their own evaluation about the applicant.

This job interview context employed in our vignette is suitable to operationalize important elements of trust violation and response processes for several reasons. First, although it does not describe a context of an established relationship, but instead an emergent one, with no prior relationship between the parties, extant research (Berg, Dickhaut, & McCabe, 1995; McKnight et al., 1998) argues that individuals may display surprisingly high initial levels of trust in strangers, even when there is no convincing evidence to support that belief. People often believe others can be considered trustworthy until proven otherwise. Second, other studies showed that the mere suspicious of committing an infraction may be sufficient to violate trust, since people are prone to believe allegations of wrongdoing even when there is no substantiation for the allegation of untrustworthy behavior (Bell & Loftus, 1989; Penrod & Cutler, 1995). Both assumptions (initial trust in parties with whom one has no prior interaction and damaged trust by unproven allegations of wrongdoing) have been successfully demonstrated empirically by prior studies (Ferrin et al., 2007; Kim et al., 2004, 2006), which employed the same setting as our vignette. Further, consistent with the information diagnosticity perspective and the belief formation and unacceptance perspective (Ferrin et al., 2007; Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993; Snyder & Stukas, 1999), in interview settings, not only do interviewers pay close attention to positive and negative information about others' integrity (Dipboye & Gaugler, 1993; Kacmar & Young, 1999), but they also form initial beliefs about the candidate very quickly and, once they are formed, those impressions are relatively resistant to change (Dougherty & Turban, 1999). For these reasons, previous research points how verbal responses given to questions regarding a job applicant's character are important defensive tactics to clear up interviewer concerns about the candidate and manage one's impressions more positively (Bolino et al., 2016; Lievens & Peeters, 2008; Tsai, Huang, Wu, & Lo, 2010).

It also should be noted that our vignette is not based on the reactions of the victim of the offense, but instead on the observer's responses. Still, our scenario is appropriate to study responses to trust violation, because there is evidence that third-parties can become outraged and react to injustices perpetrated against others (O'Reilly & Aquino, 2011). According to previous studies (Fehr & Fischbacher, 2004; Fehr & Gächter, 2002), even when they are not the target of a transgression, third-parties care about injustice experienced by others and are willing to punish deviant behavior with the same rigor as if they had experienced the injustice themselves. Because moral transgressions violate assumptions people make about how individuals should behave, they may trigger emotionally charged reactions from observers, who did not experience the harm themselves (Folger, 2001).

Manipulation

We created three versions of explanations provided by the candidate when questioned about the fraud incident. In all three conditions, the job applicant acknowledged having made inappropriate

accounting entries that enhanced reported earnings, but the explanation provided differed among the three conditions. For the moral justification strategy, the job applicant appealed to higher-order goals and greater good to justify the transgression and used the following explanation: "I did this to protect my colleagues. It was the only way to meet the quarterly profit targets. We had already failed to achieve the goals previously. If the targets weren't met once again, those people would certainly be dismissed. I knew it wasn't right, but my loyalty to my colleagues seemed to be more important at that time".

For the displacement of responsibility strategy, the job applicant explained the following: "I did this because my boss told me to do it. In my previous job, I didn't have much autonomy. Requests from bosses should not be questioned. That's how things worked out there. I knew it wasn't right, but as top management had made the request, I was afraid to refuse to comply with. My job was at stake".

Finally, for the distortion of consequences strategy, the candidate stated: "I did this, but the company was not harmed in any way. The incident didn't go public and we didn't lose any single investor. Therefore, there were no negative consequences resulting from the incident".

After reading the scenario, participants completed a questionnaire, which included items measuring trust repair.

Measures

As mentioned previously, trust can be understood as willingness to be vulnerable to others. Based on this definition, scholars argue that a proper measurement of trust should assess the extent to which a trustor is willing to voluntarily take risks at the hand of the trustee. For that reason, following previous studies (Ferrin et al., 2007; Kim et al., 2004, 2006), we adapted willingness to risk scale from Mayer and Davis (1999) to assess trusting intentions. Additionally, we adapted perceived integrity scale from Mayer and Davis (1999) to assess trusting beliefs.

Perceived integrity

Three items, on a 5-point Likert scale, ranging from 1 = strongly disagree to 5 = strongly agree, were used to assess perceptions of the job applicant's integrity. This scale was adapted from Mayer and Davis (1999). The items were: sound principles seem to guide the job applicant's behavior, the job applicant has a strong sense of justice and I like the job applicant's values. The Cronbach's α was .711.

Willingness to risk

Three items, on a 5-point Likert scale, ranging from 1 = strongly disagree to 5 = strongly agree, were used to assess how much participants, as a manager, would be willing to put themselves at risk to the job applicant. This scale was also adapted from Mayer and Davis (1999). The items were: "I wouldn't let the job applicant have any influence over issues that are important to me", "I would keep an eye on the job applicant" and "I would give the job applicant a task or problem that was critical to me, even if I could not monitor his actions". Two items were reverse-scored. The Cronbach's α was .685. The reliability for this scale was slightly higher than the levels reported by other studies using the same measure (Ferrin et al., 2007; Kim et al., 2004, 2006; Mayer & Davis, 1999).

Manipulation check

Participants were asked to indicate what was the explanation used by the candidate to justify the transgression: (a) he claimed he did for a greater good, targeting other's welfare; (b) he claimed he cannot be considered responsible for that decision, placing responsibility in another person; (c) he claimed there were no negative consequences of his action and no one was harmed. Of the 139 participants, 88% of them correctly answered the manipulation check.

Results

Confirmatory factor analyses (CFA) of the scales indicated a good fit and supported convergent validity for a two-factor model, which included perceived integrity and willingness to risk (r = 0.38), $\chi^2(8, N = 139) = 6.703$, Comparative Fit Index (CFI) = 1.000, Normed Fit Index (NFI) = 0.965, Tucker-Lewis Index (TLI) = 1.000, Root Mean Square Error of Approximation (RMSEA) = 0.000 (with p-value of being less than 0.05 equals to 0.771), Standardized Root Mean Square Residual (SRMR) = 0.036 and all item-factor loadings $\geq |0.584|$ (all p-values < 0.001). Although having perfect indices for CFI and TLI may look unusual, that can be explained by the fact that the χ^2 statistic value for this model is less than the degrees of freedom. Discriminant analyses (Bagozzi & Phillips, 1982) pointed out that the hypothesized two-factor model fits the data significantly better than the more parsimonious model (one-factor model): $\chi^2(9, N = 139) = 40.950$, CFI = 0.821, NFI = 0.789, TLI = 0.702, RMSEA = 0.160 (p-value of being less than 0.05 equals to 0.000), SRMR = 0.086. Therefore, $\Delta\ \chi^2(1, N = 139) = 34.247$, $p < 0.001$.

Table 1 presents means, standard deviations and standard errors for trusting beliefs and trusting intentions across the three explanation conditions (moral justification, displacement of responsibility and distortion of consequences). Between trusting intentions (Willingness to Risk) and trusting beliefs (Perceived Integrity), there is a positive correlation of r = 0.410 ($p<0.001$).

Table 1

Means, Standard Deviations, and Standard Errors by Condition

	Perceived Integrity			Willingness to Risk			
Type of Explanation	Mean	SD	SE	Mean	SD	SE	N
Moral Justification	2.80	0.70	0.10	2.21	0.70	0.10	47
Displacement of Responsibility	2.44	0.76	0.11	2.42	0.71	0.10	48
Distortion of Consequences	2.42	0.90	0.14	2.05	0.68	0.10	44
Full Sample	2.56	0.80	0.07	2.23	0.71	0.06	139

To assess how explanations based on different moral disengagement strategies would affect trusting intentions, we first conducted a one-way ANOVA, which revealed that there is a significant effect on type of explanations on willingness to risk, $F(2, 138) = 3.147$, $p = 0.046$, $\eta^2 = 0.044$.

To evaluate hypotheses, we then conducted finer-grained comparisons to assess how explanations based on moral justification, displacement of responsibility and distortion of consequences would compare with one another with two planned contrasts. To test Hypothesis 1a, which stated that moral justification (MJ) and displacement of responsibility (DR) strategies would lead to higher trusting intentions than distortion of consequences (DC) strategy, we conducted the first contrast. Contrast #1 coded DC as -2, and MJ and DR as 1 each. This contrast revealed that explanations based on moral justification ($M_{MJ} = 2.21$, SD = 0.70) and displacement of responsibility ($M_{DR} = 2.42$, SD = 0.71) resulted in higher willingness to risk than explanations based on distortion of consequences ($M_{DC} = 2.05$, SD = 0.68), $t(138) = 3.382$, $p = 0.042$, $d = 0.727$, supporting Hypothesis 1a.

To test Hypothesis 1b, which predicted that moral justification would lead to higher trusting intentions than displacement of responsibility, we conducted the second contrast. In contrast #2, we coded DC as 0 and MJ and DR as 1 and -1, respectively. Analysis of this contrast revealed no difference in trusting intentions between explanations based on moral justification and displacement of responsibility, $t(94) = -0.279$, $p = 0.157$. Hence, there is no support to Hypothesis 1b.

Next, to test Hypothesis 2, which predicted that trusting beliefs would mediate the relationship between type of explanation and trusting intentions, we conducted a simple mediation analysis using ordinary least squares path analysis. As our mediation hypothesis involves a multicategorical independent variable, we used an indicator coding, as developed by Hayes and Preacher (2014). To dummy-code our three categories, two dummy variables were constructed, with the variable set to 1 if a case is in its group, and 0 otherwise. Moral justification (MJ) was not explicitly coded as it was used as our reference category in the analysis. Then, using Hayes' PROCESS macro (Model 4), we assessed whether, across all explanations, perceived integrity mediated participant's willingness to risk.

Results of the mediation analysis can be seen in Figure 1 and Table 2. The relative indirect effect of a predictor multicategorical variable (D_i, as *i* denoting each group) on an outcome variable (Y) through a mediator (M) was constructed by multiplying each a_i ($D_i \rightarrow M$) by b ($M \rightarrow Y$). Specifically in our model, each effect a_i corresponds to the mean differences in perceived integrity between the different types of explanation (displacement of responsibility and distortion of consequences) relative to the control condition group (moral justification). As we coded the groups as orthogonal dummies, we excluded the problem of including simultaneously correlated variables in the same model (which may be a risk if a contrast-code was chosen, in analogous situation that multiple correlated continuous predictors were considered).

Table 2

Results of Mediation Analysis

Outcome Variables	Coeff.	Std. Err.	t	p-value	95% CI	
					Lower	Upper
Direct Effect on Perceived Integrity (R^2 = 0.048)						
Constant (Moral Justification)	2.801	0.115	24.400	0.000	2.574	3.029
Displacement of Responsibility	-0.364	0.162	-2.253	0.026	-0.683	-0.045
Distortion of Consequences	-0.377	0.165	-2.284	0.024	-0.704	-0.051
Direct Effect on Willingness to Risk (R^2 = 0.224)						
Constant (Moral Justification)	1.137	0.214	5.326	0.000	0.715	1.559
Perceived Integrity	0.384	0.069	5.586	0.000	0.248	0.520
Displacement of Responsibility	0.344	0.132	2.605	0.010	0.083	0.605
Distortion of Consequences	-0.015	0.135	-0.110	0.912	-0.282	0.252
Indirect Effect on Willingness to Risk						
Displacement of Responsibility	-0.140	0.0616[b]	[a]	[a]	-0.270	-0.028
Distortion of Consequences	-0.145	0.0723[b]	[a]	[a]	-0.301	-0.015

Note. [a]Statistics not applicable when assessing indirect effect (Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879-891. https://doi.org/10.3758/BRM.40.3.879). [b]Using 10,000 boostrap samples

We found that participants perceived transgressors who provided explanations based on displacement of responsibility and distortion of consequences to possess lower integrity than transgressors who provided explanations based on moral justification (a_{DR} = -0.364, a_{DC} = -0.377, respectively). Moreover, participants who perceived transgressors as having greater integrity were more willing to risk themselves toward the transgressor (b = 0.384). A bias-corrected bootstrap 95% confidence interval for the indirect effect of perceived integrity on Willingness to Risk based on 10,000 bootstraps samples was entirely different from zero, both for displacement of responsibility (-0.2696 to -0.0280, with point estimate ab = -0.1397), as well as for distortion of consequences (-0.3011 to -0.0147, with point estimate ab = -0.1448). Thus, we have support for Hypothesis 2. We also noted that when we

controlled for the influence of perceived integrity, residual explanations based on displacement of responsibility led to greater willingness to risk than allegations based on moral justification (c' = 0.3436), but there was no evidence that this last explanation yields difference in willingness to risk in comparison with explanations based on distortion of consequences (c' = -0.0149). Taken together, these results show that type of explanation indirectly affects willingness to risk through its effect on beliefs about others' perceived integrity.

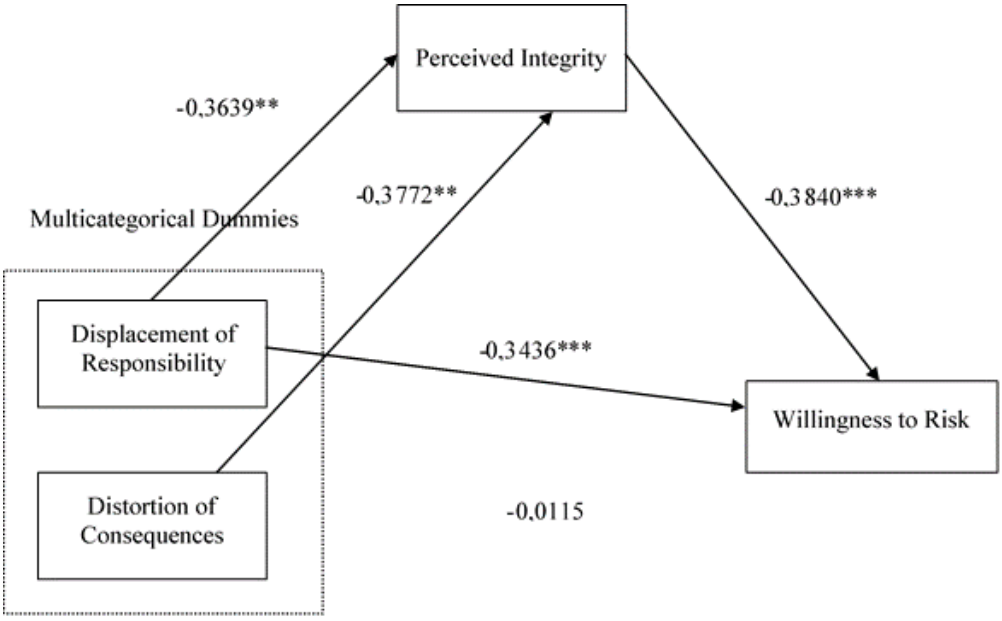

Figure 1. The Mediation Model in Path Diagram for Type of Explanation and Willingness to Risk
The group non-explicitly coded as dummy indicator is Moral Justification.

Discussion and Implications

The purpose of this study was to compare the effectiveness of explanations based on different moral disengagement on trust repair. To explore this objective, we conducted an experiment in which we manipulated explanations based on three moral disengagement strategies following a trust violation and compare how observers react to those verbal allegations.

Consistent with our hypothesis, our results showed that disregarding consequences of a transgression led to lower trust in comparison to the use of displacement of responsibility or moral justification strategies. Such explanation was inferior compared either to displacing responsibility for the transgression to authority figures or to appealing to higher-order goals resulting from values or social norms from the context. Moreover, the effects of type of explanations on trusting intentions are mediated by trusting beliefs toward the transgressor.

However, contrary to our hypothesis, moral justification did not lead to higher trust than displacement of responsibility. One possible explanation for that unexpected result relates to the items used to measure willingness to risk (trusting intentions) and the content of the displacement of responsibility strategy. Those items asked respondents, in their roles as managers, how willing they were to take risks at the hands of the applicant, once he was hired. All items were related to the level of control that managers (respondent's role) would like to have over the candidate. On the other hand, the displacement of responsibility strategy explained the wrongdoing by justifying it as obedience to authority (in our vignette, the applicant explained the misbehavior by attributing it to the management's request). Therefore, respondents could have understood such explanation as evidence that, once hired,

the candidate would be willing to conform to the manager's order. Consequently, it would be less necessary to monitor him, because he would do whatever the manager tells him to.

Such findings offer several contributions. First, we extend the literature on moral disengagement. We depart from previous studies that usually depicted moral disengagement mechanisms as internal cognitive processes to justify one's own misdeeds (Detert et al., 2008; Moore et al., 2012; Moore, 2015) by expanding their use as possible verbal statements to explain misbehavior for others. As social accounts, they not only shape observer's interpretation of a trust violation, but they also alter other's perceptions of the offender's intentions, as well as provide relevant contextual information that would otherwise be unavailable. Second, we contribute to the literature on verbal allegations and impression management. By including moral disengagement language within social accounts, not only do we draw attention to other types of explanations with ethically oriented language that has been underexplored in the literature (Dang et al., 2017), but we also show which explanations using moral disengagement language are more compelling after a trust violation. Third, we also add to the literature on trust repair by highlighting the importance of trusting beliefs as an important causal mechanism in the effects of explanations using moral disengagement language on trust repair.

Our findings also provide practical and managerial guidance for those attempting to respond to trust breaches. Considering that trust is fundamental in most workplace interactions, verbal responses following negative events between parties are important ways to solve workplace conflicts and repair relationships threatened by actual or alleged trust violations within organizations. When used wisely by employees, they may help to protect one's image from being damaged by a transgression. Therefore, verbal responses may function as effective defensive tactics to enhance one's image at work, which, in turn, is related to various important outcomes to one's career, such as hiring decisions, promotions and performance evaluations (Bolino, Kacmar, Turnley, & Gilstrap, 2008). For those reasons, in the aftermath of an integrity-based trust violation, individuals should be careful when choosing which defensive tactic to employ, because an inappropriate choice might be less effective in repairing trust that has already been impaired by a violation. Our findings suggest that, in such circumstances, verbal statements, which just attempt to minimize the negative consequences of the misbehavior, are less effective at eliciting trust than those which claim that situational circumstances had played a role in the behavior by attributing the detrimental conduct to social pressures or the dictates of others (displacement of responsibility) or by justifying the action as serving a greater good (moral justifications).

Limitations and Directions for Future Research

Some limitations have to be addressed. First, our sample was composed partially of university undergraduates. Although some researchers have expressed concern about using undergraduate samples, prior research using a similar manipulation employed in this study has found no differences in responses between undergraduate and graduate students (Kim et al., 2004). Further, we include recently graduated participants to enhance the generalizability of the results (i.e. increased breadth of working experience). Nevertheless, it would be worthwhile to replicate this study with a different population (e.g. middle- and senior-level managers). Second, we examined trust repair in a scenario-based experiment rather than in a field setting and, as a result, we face some common limitations of any experimental design. For instance, participants faced fewer costs than those that might have been associated with actual assessment and hiring decisions. Further, we assessed participant's perceptions of how trust would have been repaired rather than actual repair. Finally, although we manipulated our predictor variable (types of moral disengagement strategies) experimentally, we measured mediator and dependent variables with the same method, which might cause a bias in estimates of the relationships between them (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). We ran a confirmatory factor analysis to assess to what extent common method variance might have been a problem. Results ruled out the possibility that a single factor could account for all of the variance in our data (which would indicate that a substantial amount of common method variance is present). Nonetheless, that is still a limitation and it would be useful to implement procedural remedies to mitigate the problem in an extension of this study. Future work could

separate the measurement of the mediator and the dependent variable, for instance (Podsakoff, MacKenzie, & Podsakoff, 2012).

Despite these limitations, there are also opportunities for future research based on our findings. Our work focuses on investigating the effects of different explanations on observer's responses to a trust violation. Although prior research has found that trust can be damaged even for those who weren't directly harmed for the wrongdoing (Kim et al., 2004, 2009), one might speculate that observers may react differently from victims. Affective and intentional responses such as anger and willingness to forgive may be more intense for victims than for observers. For that reason, it would be useful to compare victims and observers' reactions to explanations based on moral disengagement strategies.

Another interesting question for future research is how individual differences among trustors may impact their reactions to explanations. Previous research has already shown how individuals may differ in their responses to verbal allegations according to some individual attributes, such as self-construal (Fehr & Gelfand, 2010) and disposition to trust (Kim et al., 2009; Mayer et al., 1995). For instance, Dang et al. (2017) found that one's tendency to disengage from moral standards (moral disengagement propensity) moderated one's reactions to social accounts using moral disengagement language in the aftermath of a transgression. Thus, in future studies, it would be useful to measure how individual differences in value priorities (Schwartz, 1992) may influence one's responses to explanations based on moral disengagement strategies. Previous empirical studies have already linked values to behaviors (Bardi & Schwartz, 2003). Therefore, as people hold various values with different degrees of importance (e.g., benevolence, conformity, self-direction), one's priority of values may influence how one may perceive and react to others' verbal responses in the aftermath of a trust violation.

Contributions

1st author: definition of manuscript scope; experiment design; data gathering and analysis; discussion of results and writing the manuscript; final revision of the paper.
2nd author: data gathering; statistical modelling and data analysis, discussion of results and writing of the manuscript; final revision of the paper.
3rd author: definition of manuscript scope; data gathering and discussion of results.

References

Ajzen, I. (2012). The theory of planned behavior. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 438-459). Los Angeles: Sage. https://doi.org/10.4135/9781446249215.n22

Anand, V., Ashforth, B. E., & Joshi, M. (2004). Business as usual: The acceptance and perpetuation of corruption in organizations. *Academy of Management Executive*, *18*(2), 39-53. https://doi.org/10.5465/AME.2004.13837437

Bagozzi, R. P., & Phillips, L. W. (1982). Representing and testing organizational theories: A holistic construal. *Administrative Science Quarterly*, *27*(3), 459-489. https://doi.org/10.2307/2392322

Bandura, A. (1986). *Foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1996). Failures in self-regulation: Energy depletion or selective disengagement? *Psychological Inquiry, 7*(1), 20-24. https://doi.org/10.1207/s15327965pli0701_3

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, *3*(3), 193-209. https://doi.org/10.1207/s15327957pspr0303_3

Bandura, A. (2011). Social cognitive theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology: Volume 1* (pp. 349-373). London: SAGE Publications Ltd.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, *71*(2), 364-374. https://doi.org/10.1037/0022-3514.71.2.364

Bardi, A., & Schwartz, S. H. (2003). Values and behavior: Strength and structure of relations. *Personality and Social Psychology Bulletin*, *29*(10), 1207-1220. https://doi.org/10.1177/0146167203254602

Barsky, A. (2011). Investigating the effects of moral disengagement and participation on unethical work behavior. *Journal of Business Ethics*, *104*(1), 59-75. https://doi.org/10.1007/s10551-011-0889-7

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin*, *91*(1), 3-26. https://doi.org/10.1037/0033-2909.91.1.3

Bell, B. E., & Loftus, E. F. (1989). Trivial persuasion in the courtroom: The power of (a few) minor details. *Journal of Personality and Social Psychology*, *56*(5), 669-679. https://doi.org/10.1037/0022-3514.56.5.669

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122-142. https://doi.org/10.1006/game.1995.1027

Bolino, M. C., Kacmar, K. M., Turnley, W. H., & Gilstrap, J. B. (2008). A multi-level review of impression management motives and behaviors. *Journal of Management*, *34*(6), 1080-1109. https://doi.org/10.1177/0149206308324325

Bolino, M., Long, D., & Turnley, W. (2016). Impression management in organizations: Critical questions, answers, and areas for future research. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*(1), 377-406. https://doi.org/10.1146/annurev-orgpsych-041015-062337

Croson, R., Boles, T., & Murnighan, J. K. (2003). Cheap talk in bargaining experiments: Lying and threats in ultimatum games. *Journal of Economic Behavior & Organization*, *51*(2), 143-159. https://doi.org/10.1016/S0167-2681(02)00092-6

Dang, C. T., Umphress, E. E., & Mitchell, M. S. (2017). Leader social accounts of subordinates' unethical behavior: Examining observer reactions to leader social accounts with moral disengagement language. *Journal of Applied Psychology*, *102*(10), 1448-1461. https://doi.org/10.1037/apl0000233

Detert, J. R., Treviño, L. K., & Sweitzer, V. L. (2008). Moral disengagement in ethical decision making: A study of antecedents and outcomes. *Journal of Applied Psychology*, *93*(2), 374-391. https://doi.org/10.1037/0021-9010.93.2.374

Dipboye, R. L., & Gaugler, B. B. (1993). Cognitive and behavioral processes in the selection interview. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 135-170). San Francisco: Jossey-Bass.

Dirks, K. T., & Ferrin, D. L. (2001). The role of trust in organizational settings. *Organization Science*, *12*(4), 450-467. https://doi.org/10.1287/orsc.12.4.450.10640

Dirks, K. T., Lewicki, R. J., & Zaheer, A. (2009). Reparing relationships within and between organizations: Building a conceptual foundation. *Academy of Management Review*, *34*(1), 68-84. https://doi.org/10.5465/amr.2009.35713285

Dougherty, T. W., & Turban, D. B. (1999). Behavioral confirmation of interviewer expectations. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 217-228). Thousand Oaks, CA: Sage.

Elangovan, A. R., Auer-Rizzi, W., & Szabo, E. (2007). Why don't I trust you now? An attributional approach to erosion of trust. *Journal of Managerial Psychology*, *22*(1), 4-24. https://doi.org/10.1108/02683940710721910

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723-733. https://doi.org/10.1287/mnsc.1110.1449

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms [Zurich IEER Working Paper No. 106]. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.495443

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-40. https://doi.org/10.1038/415137a

Fehr, R., & Gelfand, M. J. (2010). When apologies work: How matching apology components to victims' self-construals facilitates forgiveness. *Organizational Behavior and Human Decision Processes*, *113*(1), 37-50. https://doi.org/10.1016/j.obhdp.2010.04.002

Ferrin, D. L., Kim, P. H., Cooper, C. D., & Dirks, K. T. (2007). Silence speaks volumes: The effectiveness of reticence in comparison to apology and denial for responding to integrity- and competence-based trust violations. *Journal of Applied Psychology*, *92*(4), 893-908. https://doi.org/10.1037/0021-9010.92.4.893

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Folger, R. (2001). Fairness as deonance. In S. W. Gilliland, D. D. Steiner, & D. P. Skarlicki (Eds.), *Research in social issues in management* (pp. 3-33). Charlotte, NC: Information Age.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601-613. https://doi.org/10.1037/0022-3514.59.4.601

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*(2), 221-233. https://doi.org/10.1037//0022-3514.65.2.221

Gino, F., Ayal, S., & Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior & Organization*, *93*, 285-292. https://doi.org/10.1016/j.jebo.2013.04.005

Gino, F., & Galinsky, A. D. (2012). Vicarious dishonesty: When psychological closeness creates distance from one's moral compass. *Organizational Behavior and Human Decision Processes*, *119*(1), 15-26. https://doi.org/10.1016/j.obhdp.2012.03.011

Gino, F., & Pierce, L. (2009). Robin Hood under the hood: Wealth-based discrimination in illicit customer help. *Organization Science*, *21*(6), 1176-1194. https://doi.org/10.2139/ssrn.1157083

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 451-470. https://doi.org/10.1111/bmsp.12028

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Kacmar, K. M., & Young, A. M. (1999). How indirect unfavorable information is evaluated. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 229-242). Thousand Oaks, CA: Sage.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (pp. 192-238). Lincoln: University of Nebraska Press.

Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, *120*(1), 1-14. https://doi.org/10.1016/j.obhdp.2012.08.004

Kim, P. H., Diekmann, K. A., & Tenbrunsel, A. E. (2003). Flattery may get you somewhere: The strategic implications of providing positive vs. negative feedback about ability vs. ethicality in negotiation. *Organizational Behavior and Human Decision Processes*, *90*(2), 225-243. https://doi.org/10.1016/S0749-5978(02)00522-8

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review*, *34*(3), 401-422. https://doi.org/10.5465/AMR.2009.40631887

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, *99*(1), 49-65. https://doi.org/10.1016/j.obhdp.2005.07.002

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology*, *89*(1), 104-118. https://doi.org/10.1037/0021-9010.89.1.104

Kim, P. H., & Harmon, D. J. (2014). Justifying one's transgressions: How rationalizations based on equity, equality, and need affect trust after its violation. *Journal of Experimental Psychology: Applied*, *20*(4), 365-379. https://doi.org/10.1037/xap0000030

Kish-Gephart, J., Detert, J., Treviño, L. K., Baker, V., & Martin, S. (2014). Situational moral disengagement: Can the effects of self-interest be mitigated? *Journal of Business Ethics*, *125*(2), 267-285. https://doi.org/10.1007/s10551-013-1909-6

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, *53*, 107-117. https://doi.org/10.1016/j.jesp.2014.03.005

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88-106. https://doi.org/10.1016/j.obhdp.2014.10.007

Lewicki, R. J., & Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*(1), 287-313. https://doi.org/10.1146/annurev-orgpsych-032516-113147

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management*, *32*(6), 991-1022. https://doi.org/10.1177/0149206306294405

Lievens, F., & Peeters, H. (2008). Interviewers' sensitivity to impression management tactics in structured interviews. *European Journal of Psychological Assessment*, *24*(3), 174-180. https://doi.org/10.1027/1015-5759.24.3.174

Martin, S. R., Kish-Gephart, J. J., & Detert, J. R. (2014). Blind forces: Ethical infrastructures and moral disengagement in organizations. *Organizational Psychology Review*, *4*(4), 295-325. https://doi.org/10.1177/2041386613518576

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, *84*(1), 123-136. https://doi.org/10.1037/0021-9010.84.1.123

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709-734. https://doi.org/10.2307/258792

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review*, *23*(3), 473-490. https://doi.org/10.5465/AMR.1998.926622

Moore, C. (2015). Moral disengagement. *Current Opinion in Psychology*, *6*, 199-204. https://doi.org/10.1016/j.copsyc.2015.07.018

Moore, C., Detert, J. R., Treviño, L. K., Baker, V. L., & Mayer, D. M. (2012). Why employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychology*, *65*(1), 1-48. https://doi.org/10.1111/j.1744-6570.2011.01237.x

O'Reilly, J., & Aquino, K. (2011). A model of third parties' morally motivated responses to mistreatment in organizations. *Academy of Management Review*, *36*(3), 526-543. https://doi.org/10.5465/AMR.2011.61031810

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*(4), 817-845. https://doi.org/10.1037/1076-8971.1.4.817

Peus, C., Wesche, J. S., Streicher, B., Braun, S., & Frey, D. (2012). Authentic leadership: An empirical test of its antecedents, consequences, and mediating mechanisms. *Journal of Business Ethics*, *107*(3), 331-348. https://doi.org/10.1007/s10551-011-1042-3

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*(1), 539-569. https://doi.org/10.1146/annurev-psych-120710-100452

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879-891. https://doi.org/10.3758/BRM.40.3.879

Reynolds, S. J., Dang, C. T., Yam, K. C., & Leavitt, K. (2014). The role of moral knowledge in everyday immorality: What does it matter if I know what is right? *Organizational Behavior and Human Decision Processes*, *123*(2), 124-137. https://doi.org/10.1016/j.obhdp.2013.10.008

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393-404. https://doi.org/10.5465/AMR.1998.926617

Schlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations*. Monterey, CA: Brooks/Cole.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology, 25*, 1-65. https://doi.org/10.1016/S0065-2601(08)60281-6

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1-19. https://doi.org/10.1016/j.obhdp.2006.05.005

Shapiro, D. L. (1991). The effects of explanations on negative reactions to deceit. *Administrative Science Quarterly*, *36*(4), 614-630. https://doi.org/10.2307/2393276

Shaver, K. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer-Verlag.

Shaw, J. C., Wild, E., & Colquitt, J. A. (2003). To justify or excuse?: A meta-analytic review of the effects of explanations. *Journal of Applied Psychology*, *88*(3), 444-458. https://doi.org/10.1037/0021-9010.88.3.444

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin*, *37*(3), 330-349. https://doi.org/10.1177/0146167211398138

Snyder, M., & Stukas, A. A. (1999). Interpersonal processes: The interplay of cognitive, motivational, and behavioral activities in social interaction. *Annual Review of Psychology*, *50*(1), 273-303. https://doi.org/10.1146/annurev.psych.50.1.273

Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research*, *17*(2), 223-236. https://doi.org/10.1023/B:SORE.0000027411.35832.53

Tomlinson, E. C., Dineen, B. R., & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management*, *30*(2), 165-187. https://doi.org/10.1016/j.jm.2003.01.003

Tomlinson, E. C., & Mayer, R. C. (2009). The role of causal attribution dimensions In trust repair. *Academy of Management Review*, *34*(1), 85-104. https://doi.org/10.5465/AMR.2009.35713291

Tsai, W.-C., Huang, T.-C., Wu, C.-Y., & Lo, I.-H. (2010). Disentangling the effects of applicant defensive impression management tactics in job interviews. *International Journal of Selection and Assessment*, *18*(2), 131-140. https://doi.org/10.1111/j.1468-2389.2010.00495.x

Tsang, J.-A. (2002). Moral rationalization and the integration of situational factors and psychological processes in immoral behavior. *Review of General Psychology*, *6*(1), 25-50. https://doi.org/10.1037//1089-2680.6.1.25

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548-573. https://doi.org/10.1037/0033-295X.92.4.548

Weiner, B. (1986). *An attributional model of motivation and emotion*. New York: Springer-Verlag.

Weiner, B., Amirkhan, J., Folkes, V. S., & Verette, J. A. (1987). An attributional analysis of excuse giving: Studies of a naive theory of emotion. *Journal of Personality and Social Psychology*, *52*(2), 316-324. https://doi.org/10.1037/0022-3514.52.2.316

Wood, R. E., & Mitchell, T. R. (1981). Manager behavior in a social context: The impact of impression management on attributions and disciplinary actions. *Organizational Behavior and Human Performance*, *28*(3), 356-378. https://doi.org/10.1016/0030-5073(81)90004-0

Authors

Tatiana Iwai
Insper, Rua Quata, 300, 04546-042, São Paulo, SP, Brazil. E-mail address: tatianai@insper.edu.br. http://orcid.org/0000-0002-8733-5369

João Vinícius de França Carvalho
FEA/USP, Av. Prof. Luciano Gualberto, 908, Cidade Universitária, FEA 3, 05508-010, São Paulo, SP, Brazil. E-mail address: jvfcarvalho@usp.br. https://orcid.org/0000-0002-1076-662X

Victor Marson Lalli
Insper, Rua Quata, 300, 04546-042, São Paulo, SP, Brazil. E-mail address: v.lalli@outlook.com. https://orcid.org/0000-0002-3730-6279