When Might Forgiveness Help Solve Social Dilemmas?*

Jane Sell
Texas A&M University

Katie Constantin
Oklahoma State University

Chantrey J. Murphy
California State University, Long Beach

*Paper for presentation at the Ostrom Workshop, Indiana University, October 20, 2021.
This paper contains much of the information from our longer review of the literature in: Sell, Constantin and Murphy, (2020) "Reputation, Forgiveness, and Solving Problems of Cooperation", Thye, S.R. and Lawler, E.J. (Ed.) *Advances in Group Processes*, Emerald Publishing Limited, Bingley, pp. 109-134.

When Might Forgiveness Help Solve Social Dilemmas?

ABSTRACT

When people know that others have a reputation for being cooperative, their decisions about how and whether they should cooperate seems straightforward. However, what if people know that others have a noncooperative reputation?

How reputations develop, are assessed, and can be repaired are often complicated issues. To address some of these issues, we first develop formal definitions of observers and reputation. We then ask how noncooperative or "bad" reputations might be repaired. When the goal is solving a social dilemma, forgiveness is important because it can enable social integration and cohesion. Based on the developed definitions and past research we suggest some possibilities for forgiveness and reconciliation. We also consider an experimental paradigm to investigate reputations.

How do people decide if they should cooperate with others? This general question is critical for a wide variety of interactions, ranging from small and relatively inconsequential (such as how to avoid collisions on a field track when there are a variety of activities) to large and life altering (Refugee agencies working with each other to ensure the safety of those needing help.)

When the outcome for both parties is achievable only through cooperation, as long as all actors benefit, cooperation can be sustained. When incentives for cooperation are extremely high and other alternatives are unavailable, coordination can be problematic, but intentions of actors are not. However, there are other contexts that *require* cooperation to develop solutions but actors' incentives to cooperate are mixed. One class of such contexts are social dilemmas and solving these require instituting mechanisms to ensure cooperation. The particular mechanisms that we address are reputation and associated with reputation, forgiveness.

SOCIAL DILEMMAS

All social dilemmas share the defining characteristic of a conflict between the best response for an individual, or a set of individuals, and the best response for all others in the group (for definitions of social dilemmas, see Dawes, 1980; Kollock, 1998).

There is substantial literature on possible solutions to social dilemmas that spans many disciplines. Many of the solutions focus on the material incentive structures that characterize the context. For example, if a community group monitors the behavior of its members and punishes those who do not cooperate to the same extent as others, cooperation in that context is encouraged, and free riding is discouraged. (See Fehr & Gächter, 2000; Olson, 1965; Ostrom, 1990; 1998; Simpson & Willer, 2015.)

There are other kinds of solutions that do not directly manipulate the readily apparent material incentives of the setting. It is these kinds of solutions we address, those that stress the

relationships among and between individual actors. These relationships can be in the form of beliefs about oneself and others, or group identity for example. In this paper, we focus on the *reputation* of an actor, where an actor can be any identifiable entity. Such an entity could be an individual, an organization, a nation-state, etc.

Because social dilemmas are so pervasive, they are topics of interest for many different disciplines. And, because reputation is often at the heart of social dilemmas, it has also been of interest across different disciplines. However, those disciplines also have varying definitions, conceptualizations, and effects of reputation. We briefly discuss different conceptualizations and then offer a way to combine them comprehensively. We argue that the definitions themselves provide insight into how reputations are both constructed and possibly reconstructed. In particular, we focus on how negative reputations for cooperation might be repaired.

What is a reputation?

Definitions of reputation across different fields often differ based on the unit of analysis. We discuss very generally, ideas of reputation of an actor, where an actor can be a person, a group, an organization, a state, etc.

The Actor as an Organization, Firm or State

In many fields the organization itself is the actor. Reputation includes beliefs about what to expect from the organization in the future and favorability (Lange, Lee, & Dai, 2011). In investigations of firms, McDonnell and King (2018) argue that reputations are viewed as shared perceptions that frequently rely on perceptions of past behavior within a particular domain (Fombrun & Shanley, 1990; King & Whetten, 2008; Roberts & Dowling, 2002; Sorenson, 2014).

The area of public relations has evolved to the point of describing itself as "the discipline which looks after reputation, with the aim of earning understanding and support and influencing

opinion and behavior" (Chartered Institute of Public Relations, 2019). As such, reputations are managed based on the assessment of external stakeholders who are invested in the success of a company or organization (Cağin Bektaş, 2018; Gibson, Gonzales, & Castanon, 2006; Kottasz & Bennett, 2016).

Ordinarily, we think of reputations as operating in the present, but because reputations involve the idea of shared perceptions or narratives, it is possible that reputation can function and "change" the past as well. Within the sociology of art and history, for example, reputation has been defined as an aspect of collective memory or a living image of the past (Halbwachs, 1980), a "collective definition" based on what is known about an individual or group (Lang & Lang, 1988, p. 84), and embellishments of the public's memory (Lewis, 1975). According to Lang and Lang (1988), there are two components of reputations: recognition and renown. Recognition refers to the amount of esteem the actor or actors hold, and renown refers to how well the actor is known beyond a particular inner circle. Reputations that survive the test of time are those that are maintained by institutions that have the ability to archive historical evidence supporting specific reputational claims (Fine, 1996; 2019; Lang & Lang, 1988). Related to this, there is a literature on states and their reputation for resolve.

Reputations for resolve are believed to be highly valued characteristics for all states because they reflect others' perceptions about a state's willingness to engage in military disputes (Weisiger & Yarhi-Milo, 2015; Wiegand, 2011). Reputations within this body of work are viewed primarily as a product of the state's past actions, and as a result, allows them to be fortified and invested in further actions (Guisinger & Smith, 2002; Klabi, Mellouli, & Rekik, 2014; Weisiger & Yarhi-Milo, 2015).Reputations for resolve force others to respect the

credibility of one's threats and thereby increase the likelihood that those threats will be believed during future altercations (Gibler, 2008; Jervis, 1988; Snyder, 1984).

As Satori (2005) notes, if a state makes a claim and then backs down or withdraws that claim, they will develop a reputation for dishonesty or bluffing. These types of negative reputations can hinder a state's credibility and thereby hinder its capability to attain its goals for future international relations (Gibler, 2008). On the other hand, actually acting on a threat through coercion can be costly, and as a consequence, sometimes states do not fight for their reputations, and instead acquiesce. As discussed, and empirically illustrated by Sechser, 2018, a reputation for resolve is only worth defending if it is going to have future implications. If the other actor involved in the crisis is unlikely to be a rival in the future (either through geographical space, past aggression, and military power), then defending the reputation for resolve may be less important.

The Actor as an Individual or Group

Much of the psychological literature cites Emler's (1990) definition of reputation for an individual: "a set of judgments a community makes about the personal qualities of one of its members" (p. 171). When examining individual level phenomena, reputation serves as an instrument for (pro)social order (Darby & Schlenker, 1989; Emler, St. James, & Faucheux, 2013; Jazaieri, Logli, Allison, Campos, Young, & Keltner, 2018; Simpson & Willer, 2015;). an individual actor's ability to fulfill their wants and needs is predicated on their reputation, which they can then use to cultivate social ties and exchange opportunities.

The ability to communicate with others and share direct firsthand accounts, or share other's indirect secondhand accounts, is what facilitates reputations of community (Anderson & Shirako, 2008; Bohnet & Huck, 2004; Emler et al., 2013). It is through narrative accounts that

we are able to establish the traits and personality of a person and how well they meet expectations of how they should act, and whether they are a good candidate for future social exchanges. However, the type, context and source of information contributing to a reputation varies widely which also suggests that validity of reputations vary. This, in part, is due to the difficulty with managing information about all members of a community. As an example, reputations tend to be more directly linked to a person's behavioral history when the person is also someone of status (e.g., leader). This is because fewer resources are necessary when many community members retain information on only a few others, particularly those who are prominent members. As such, retaining firsthand and secondhand information about the past behavior of a select few individuals in the community yields reputations that are perceived as valid for those persons.

Much of the literature investigating reputations describe them as social instruments that alert community members about one another's history of cooperation and other general qualities. With the innumerable exchange opportunities and exchange partners available, and with the threat of engaging with those who would take advantage of our resources for their own selfish gains, reputations allow us to navigate society and identify those whom we prefer to limit engagement. As such, people are also motivated to maintain "good" reputations that signal to others that they can be trusted to act in ways that do not negatively impact others in the community (Darby & Schlenker, 1989; Griskevicius, Tybur, & Van den Bergh, 2010; Sallot, 1993).

Game Theory Models (using both groups, individuals, organizations as actors)

The game theory literature often overlaps with evolutionary perspectives and views reputation as a critical component in decision making. For example, Haidt (2007) argues that

combining evolutionary concepts with Durkheimian insights explains the development of communities with shared norms which encourages the development of reputations for cooperation. In this way, morality develops to bind societies to better survive as a group.

Many of these models depend upon reputations of past helping. Sugden (1986) developed an argument based on "good standing" which is a reputation based on helping. An actor loses good standing if they do not extend help. Therefore, the existence of this good standing, and the potential to lose it, acts as an insurance principle to promote cooperation.

Nowak and Sigmund (1998) analyzed models of populations of actors who could help or not help a partner. In their models, each actor had an image score, which is based on their decisions to either help or not help another (the cost for the donor is assumed to be less than the benefit that a recipient receives). However, if the donor does not help, no one receives a payoff. This is conceptualized in terms of incremental fitness or reproduction in evolutionary models. Nowak and Sigmund (2005) and Suzuki and Akiyama (2005) later expanded these models to consider larger groups in which pairs of actors might only meet a couple of times. Again, reputation in the sense of an image score plays the pivotal role in the development of cooperation.

COMMONALITIES AND DIFFERENCES AMONG DEFINITIONS: FORMALIZING DEFINITIONS

Although there are many differences among the definitions used in different approaches, we offer the following formal definitions that attempt to distill the major elements across different fields and do justice to the concerns. Our purpose is to develop exact class concepts that enable identification of reputation and then theoretical extension and application based on those definitions (Sell 2018).

We begin with the idea of a domain. A domain is a defined context that contains the elements below.

For a given Domain, D, at least 4 sets can exist:

Set of Actors, A, = $\{a_i\}$; i= (1,2,….,N). An actor a_i is an individual, group, organization or other unit.

Set of Behaviors, B, = $\{b_j\}$; j = (1, 2,….,N). An element e_i is a behavior b_j if and only if:

a. At time t_1, each actor a_i acts with some degree of independence from another actor or actors in the performance of each e_j;

b. Each performance is a choice between at least one e_i and one e_j.

c. At some point or points in time, t_1, t_2, …., t_n, a_i performs at least one e_j.

The behaviors performed by an actor have only two constraints. First, the actor must act with some degree of independence from other actors (if two actors act together, they are considered as a single actor). Secondly, there is a choice involved between at least two behaviors. It is possible that actors together might develop a reputation as a unit (for example, the authors of an article might develop a reputation based upon their jointly produced article), but the authors might also be viewed as separate actors if observers see them as such. The practice of some journals to separate out the responsibilities of the different authors might be viewed as an attempt to separate out reputations.

Set of Observers, O, = $\{o_k\}$; k = (1, 2,….,N). An actor is an observer between an actor, a_i, and a behavior, b_j, if and only if at each point in time, the actor connects a_i with b_j such that the b_j is implied by a_i with probability p.

Observers are those who connect an actor with a particular behavior such that the actor implies that behavior. So, for example, if Xander has acted noncooperatively in the past, another actor might develop a perception of Xander such that they connect Xander to noncooperation with a high degree of probability. In other words, they think that Xander's past behavior is viewed as a part of Xander's character.

> **Set of Reputations**, R, = $\{r_l\}$; l = (1, 2,….,N). Observers create a reputation, if and only if an observer, o_k, communicates to actors other than o_k.

The important aspect that differentiates an observer of a behavior from the creation of a reputation is that the implication between the actor and the behavior is transmitted to others. How it might be transmitted could vary greatly. It could be a tweet, face-to-face gossip, a newspaper article, or a blog entry. This particular idea of a reputation dissolves the difference sometimes made between indirect and direct experience with particular actors (mentioned within political science literature). We view this as an advantage because it separates the *way* in which behavior might be observed from the strength of the communication. Who makes the observation (and the communication) can be examined independently from the type and strength of the communication.

> **Generalizability from Reputational Domain 1 to Reputational Domain 2:**
>
> The definition of an observer implies that some generalizability among behaviors may occur. The degree of generalizability is dependent upon the observers and their connection across behaviors that might belong in two different domains. The

generalizability coefficient from Domain 1 to Domain 2 is the probability that they are dependent. The probability of $R_1 | R_2$ is equivalent to the generalizability coefficient between domain 1 and domain 2.

In the sense of the logical connection, (if x than y), behaviors in one area can affect connection in another area. So, for example, a firm's reputation for safety may (or may not) be connected to a firm's reputation for reliability by observers. Whether there is generalizability would be evidenced by observers' perceptions as characterized by the mathematical independence of one behavior from another.

The reputation history, RH, is the sum of reputations over time and over domains.

These definitions reach across the different disciplines, in part because the concept of reputation is not tied to an individual or a firm or a nation, but rather concerns any actor. Because these definitions are abstract, they can apply to a great many different contexts and disciplinary concerns. Neither the content nor the valence of the reputation is addressed, and so the definitions can apply to positive, negative, or neutral reputations. This aspect of the definitions help sort out some of the definitional problems in the literature in which reputations are only discussed in terms of "bad" or "good." Further, these definitions scope out necessary conditions for the formation of a reputation: a perceived connection between a behavior involving choices and the actor itself *and* the transmission of that perceived connection by observers. Such a definition enables consideration of those who are living as well as those who are dead and enables consideration of changing information (communicated through observers)

and therefore changing reputations. The concept of domains helps to delineate how organizations, firms, individuals and/or nation-states might have very different reputations in different settings. So for example, while a company might have a reputation for artistic, unusual clothing, they might also have a reputation for unfair labor practices. In this example, it is likely that, in our terms, the reputation in one domain does not generalize to the other domain because knowing information about one behavior does not provide information about another (that is, they are independent). At the same time, reputations can generalize from one domain to another, and this would be demonstrated by the likelihood (or probability) that knowledge of reputation in one domain provides knowledge of other domains. As an example, if a professor is observed to be verbally abusive to her children, it is likely this information would generalize to a classroom reputation. The generalizability coefficient represents how likely observers believe behavior in one situation predicts behavior in another situation. This coefficient also applies regardless of the content of the domains.

With this formalization, we now consider the following: (1) under what conditions can reputations be employed and (2) how can reputations be changed or repaired?

REPUTATION VS TRUST BASED ON REALIZATION OF SHARED INTERESTS

In some cases, reputations of actors are not important. As mentioned earlier, there are times when the context itself demands cooperation. In such contexts, even if a person has not been cooperative in the past, they will cooperate if it is obviously in their best interest and the only way a goal can be reached. (That is, it is a pure cooperation setting.) But when faced with a social dilemma, the incentive structure changes and incentives are mixed. In these settings, one of the most important aspects of cooperation relates to how long individuals expect to interact

with the same set of others. There are two well-known general sort of issues related to interaction that open different possibilities for cooperation. One of these relates to the development of a strategy with others who also know they are interacting within the same group. The other issue relates to *what* an individual actor knows about the others interaction.

While at each point in time, the properties of the social dilemma make it individually rational to defect, repeated interactions can change decisions. Without addressing individual rationality, if interactions are repeated with the same partners or group members, then cooperating (contributing or restraint from taking) is a possible strategy. A rational actor can see that if she is interacting with the group for a long period of time, and she defects early in the interaction, others will as well. If that is the case, the group and all the individuals in the group will have lower earnings than if they had cooperated, at least initially.

The formal development of this argument is generally credited to the folk theorem (so called because it was a generally understood idea, or folk knowledge), which posits a whole range of history-contingent strategies that allow for cooperation. Two conditions are necessary: (a) at some point, an actor's *potential* gain is greater than the cost of contribution and (b) the discount rate is sufficiently large such that contributing remains an individually rational strategy. (See Aumann, 1987 and Fudenberg & Maskin, 1986.) The discount rate stipulation ensures that the actor believes the future will hold some promise or at least that it will come. If survival is questionable, for example, if the family is near death, these two conditions may not be met. In essence, an actor must believe that the future within the group will be a possibility. From this point of view, social dilemmas can be solved rationally, although it is difficult to predict exactly how. The folk theorem does not rule out many possibilities. It does imply, however, that trust that others

will cooperate can develop from the recognition that others also realize the implications of the folk theorem.

With repeated decisions, there is also the opportunity for actors to learn about others or to learn about or develop reputations. In this way, reputations can be another way to solve social dilemmas. If an actor knows what behavior the other (or others) is most likely to choose, then the social dilemma can be transformed. So, for example, if an actor knows that others are cooperative, choices become easier. Of course, this whole issue of strategy can quickly become complicated because there is an incentive to appear to be cooperative so that others will cooperate and, in this way, deceive other members so that one can take advantage of their cooperation. Because everyone knows that everyone else might be "faking" being a cooperative type, reputations can be difficult to develop and maintain. Additionally, when actors know that others have a noncooperative reputation, cooperation seems doomed. The issue of reputation is also related to the broader issue of what kind of information actors have about themselves, others, and the context.

One game theoretic implication of knowing the types of actors involved in the social dilemma is that it helps transform a setting of incomplete information to one of imperfect information**.** In game theory terms, people cannot make decisions if information is incomplete—that is, where the context or events are not clear enough to enable some sort of calculation about the distribution of events. Incomplete games cannot be solved because (among other things) the payoffs of others are unknown and so actors' strategies cannot be modeled. In a series of papers, Harsanyi (1967; 1968a; 1968b) argued that a game of incomplete information could be transformed into a game of *imperfect* information. This argument is called the Harsanyi transformation and has had a powerful effect upon game theory. If the game is incomplete, we cannot solve it. If we can transform it, however, then we can solve it. We can do this either by

making assumptions about the distribution of actors of a certain type, that is, by reputation, or by assuming the distribution of the rules of the game (Binmore, 1992, p. 503). We know, from many literatures, that actors generally seek to decrease unpredictability in the sense of unknown properties. (For discussion of these different literatures, see Brewer, 2007; Kollock, 1994; Lawler, Thye & Yoon, 2000; 2009; Molm, Takahashi & Peterson, 2000; Williamson, 1981.) If the actor is presented with information about the context or reputation of others, actors can develop strategies for their best response.

Because reputation provides one way to solve social dilemmas, as we have discussed, it has been the subject of a great deal of research. As discussed in Simpson and Willer (2015), research demonstrates that those with cooperative reputations are respected more and others seek them out (Barclay, 2004; Barclay & Willer 2007; Feinberg, Willer & Schultz, 2014; Willer, 2009). Similarly, firms with more positive reputations are rewarded with profits.[i] In the game theory sense, reputation also refers to the dynamic process of developing and modifying reputation. That is, even if an actor is *not* a cooperative type, because of the rewards of having a cooperative reputation, it is beneficial to project the image of being cooperative. Of course, everyone knows these potential benefits, and so there are efforts to validate reputations.

REPUTATIONS AND GOSSIP, OSTRACISM, AND FORGIVENESS

In our definitions, observers see actors' behaviors that are the results of a choice. The observers connect an actor and the particular behavior. But reputations are not developed until they are communicated. The act of communication is often a complicated issue, however. Communication can take many different forms and occur differently at different points of time. Part of the communication can be labeled as gossip, although the way in which gossip has been

defined varies a great deal. Dunbar (2004) argues that gossip can be important to prevent rampant exploitation. Experimental studies involving social dilemmas support this general finding (Feinberg et al., 2014). In particular, participants use gossip (or communications) to structure their own choices, effectively ostracizing those who were noncooperative. But, knowing that the community can gossip also leads to more cooperation, that is, people who were initially noncooperative were apparently compelled to become more cooperative because of the consequences of gossip.

Reputations are information on a community or group level and, from a game theoretic level, help both coordinate and suggest strategies. As suggested above, positive reputations are important and sought after. But what happens when negative reputations accrue to an actor?

Ostracism is one response to a negative reputation, and the threat of this can help maintain cooperation (Feinberg et al., 2014). But it also splinters groups. Can recovery from negative reputations occur such that groups do not resort to expelling community members? While this is an important subject for the study of peace and reconciliation processes (Amstutz, 2005; Govier, 2006; Hayner, 2010), there is little social dilemma research on this topic (for an exception on the study of noncooperative types, see Eriksson & Strimling, 2012). Ostracism also brings in the issue of morality, an issue, as discussed by Simpson et al., (2017) that is surprisingly underdeveloped in many discussions of prosocial behavior. Their research investigated how feedback stressing the morality of acts impacted cooperation. In particular, they demonstrate how those who make moral judgments perceive themselves as possessing a more moral identity and being more trustworthy than others and how those who observe them also trust them more.

But, can this process be developed for those who have negative reputations? This evokes two issues: (a) the morality of the person who has transgressed and (b) the morality of those to whom the transgression has occurred. Stated in another way, can forgiveness be granted and cooperation maintained?

By our definition of *reputation* and *reputational history*, reputations are built over time. Because this is the case, reputations can change. So, by definition, reputations can be "repaired" if they become sullied. But, how does this happen? If reputations are relatively stable and positive, for example, perhaps one or two noncooperative acts can be "dismissed." Are these acts forgiven? Is the dismission or discounting of an act equivalent to forgiveness?

As would be expected, there is a long history of discussion concerning the nature of forgiveness within philosophy and psychology, especially clinical psychology. The general issues center upon what must occur for forgiveness; who has the standing to forgive; and when forgiveness is morally just (Hughes & Warmke, 2017). Forgiveness also has an enigmatic quality—it is granted for past behavior but can only occur in the present and is meant to affect the future.

The clinical psychology field seems to have developed a consensus that forgiveness is important for both parties but focuses on the forgivers rather than the forgiven. Forgiveness in such a scenario is a reduction in angry or vengeful thoughts and an increase in positive thoughts and actions (Wade, Hoyt, Kidwell, & Worthington, 2014; Wade & Worthington, 2003; Worthington, 2005). From this point of view, forgiveness may not involve forgetting or excusing. Reconciliation might occur, but it might not. It is worth noting that this idea that forgiveness gives benefits to those who forgive is also an important point in many religious views as well. In Christianity, there is admonition that one's own forgiveness, increases worth:

"For if you forgive others their trespasses, your Heavenly Father will also forgive you" (Matthew 6:14). In the Qur'an, people are instructed to forgive and find reward with God.

This makes clear that forgiveness is an interactional issue, involving the trespassers as well as those who are trespassed against. Forgiveness is not entirely controlled by those holding the reputation in which transgression has occurred, rather it is granted by the observers. These observers might be directly affected by the wrongdoing performed by an actor. For example, they may be the victims of insensitivity, bad or brutal behavior, or harm to their family. But as Radzik (2010) argues, they do not necessarily have to be direct recipients of the harm performed by the harm-doer. The necessary component is whether some degree of "moral anger" is experienced by the observer (be they victims or not). In this way, we can think of the moral community—an important concept for the consideration of social dilemmas for example.

As Haidt (2007) indicated, community and morality are important in discussions of evolutionary models, as well. These models seriously consider the importance of *reconciliation* models of forgiveness (Axelrod, 1980a, 1980b; de Waal & Pokorny, 2005). According to these views, forgiveness evolved as a mechanism for affirming mutual cooperation between agents after an act of defection. While punishment for noncooperation might be effective and sometimes required, it also necessitates relatively constant monitoring. Because monitoring is costly, some sort of reconciliation processes are selected to solve this issue (Petersen et al., 2010; de Waal, 1996). Thus, when confronted with exploitation, two distinct factors are involved when deciding consequences for the punished: (a) the exploiter's potential value in the future and (b) the magnitude of the seriousness of the exploitation (Petersen et al., 2012).

Empirical studies of forgiveness

The empirical studies of forgiveness tend to be in psychology or in interdisciplinary approaches which emphasize forgiveness at the individual level rather than within an interaction, group, or community. In part, this most likely occurs because of the emphasis on the social benefits of forgiveness as we have discussed.

Several measures of individual forgiveness have been developed. One is the 19-Item Transgression-Related Interpersonal Motivations Scale (TRIM-19; McCullough & Hoyt, 2002; McCullough et al., 1998). The TRIM contains items that the respondent indicates agreement with concerning a particular trespassing event and person. Examples of items include "I'm going to get even," and "Despite what he/she did, I want us to have a positive relationship again." Williamson, Gonzales, Fernandez, and Williams (2014) developed a measure of forgiveness aversion, and they argue that the key features that predict if actors are averse to forgiveness include whether the trespass was repeated, if rumination is involved, and if revenge is a motivation. In their discussion of their studies of forgiveness aversion, they highlight that forgiveness is often built upon empathy or altruism while forgiveness aversion is based upon fear. Example questions include such things as "I am confident that the offender won't continue to hurt me in the future if I forgive," and "By forgiving, I may appear to be weak in front of the offender and others."

As indicated in the literature, forgiveness is calibrated, in part, on the perception of the remorse of the perpetrator(s). Social processes such as apologizing and demonstrating remorse seem to offer effective ways for actors to convey that they are aware of their transgression and are eager to limit the damage. Of course, the sincerity of the apology is an important issue. As this is the case, apologies have been investigated in different domains and with different results.

In business relations, Abeler, Calaki, Andree, & Basek (2010) investigated apologies by firms. In an innovative study, they tested whether a firm's apology can influence consumer's behavior. The researchers collaborated with a firm on German eBay and first found customers who had given negative or neutral evaluations. Then, they randomly assigned each of those customers to receive either an apology, a small monetary compensation, or a large monetary compensation for the withdrawal of their negative or neutral evaluation. The apology did not include any admission of guilt, it was just a simple apology. Comparing the evaluations withdrawn by treatment, the researchers found that apologies were much more effective than either of the compensation treatments.

Ho (2012) examined the use of apologies in trust games. As does Abeler, he conceptualizes apologies as "cheap talk", meaning that they are not contractual, and since there is no enforcement mechanism, they really have no economic value. However, he argues that they are sent as signals for the future fitness for interaction. And indeed, very simple apologies without reference to guilt had positive effects in trust games, most especially if they were delivered early in the interactions.

But, can this process be developed for those who already have negative reputations? That is, those who *enter* into a context in which their (negative) reputation already exists?

In an early study, Darby and Schlenker (1989) examined the perceptions of individual school aged children who were told that their property had been damaged by a fellow classmate. They also varied the base reputation and behavior of the offending child by describing them as having a (*good*/*bad*) reputation and expressing (*remorse*/*joy*) after the event took place. As expected, offending children with good reputations before the incident were believed to be genuinely apologetic after having damaged property compared to children with bad reputations.

Furthermore, punishments for the transgression were significantly less when the offending child had a previously good reputation. Interestingly, possessing a bad reputation and showing remorse did lead to reduced punishment (compared to showing joy), but the punishment was still greater compared to the offending child with a good reputation who showed no remorse. These findings highlight the influence of existing reputations in determining social repercussions for harmful behavior.

In social dilemma settings, as in the settings most often discussed in evolutionary models, community actors are dependent upon the cooperation of others. Without any other information available, we would expect that reputation would be used to develop group interaction. In social dilemma settings, where cooperation helps the entire group, those with competitive reputations would not be trusted, and group cooperation would quickly decay. But suppose that those with competitive reputations could engage in other strategies to try to persuade others that they would change and act cooperatively. It is obviously in an actor's self-interest to persuade others that they can be cooperative. When will others be persuaded? Others are skeptical of claims if they are self-serving. So, apologies or explanations to the entire group would be important for an actor attempting to claim that they will *not* act noncooperatively, even if they had before. This would be necessary because it would function as a reason for the modification from "defector" to "cooperator" and also serve as a public commitment. Further, this explanation would need to be viewed as sincere. Additionally, and importantly, the reputations of the others in the group are implicated. Remembering the different arguments, including those in evolutionary models, "forgiveness" is an important asset in reputations. At the same time, "unconditional forgiveness" can be problematic for social dilemmas because such a reputation could signal that others could just free-ride on the forgiving group members. So, a positive reputation for a cooperative group

member is “contingent cooperator” or “contingent forgiver,” someone who can cooperate and forgive, but not unconditionally. (Also see Ostrom’s (1998) discussion of contingent cooperators.)

To use our definitions then, we can see that those with a negative reputation need to offer either different behavior or promise of a different behavior to modify their reputation. By definition, this cannot occur unless there are observers. As such, this means some public announcement or public action must occur. Also by our definitions, if *both* the reputation of those who have been noncooperative in the past and those who have been cooperative can be moved further toward “trusted cooperator,” the probability of actors assessing cooperation increases. Indeed the reputations for *both* those with positive and those with negative reputations can become more positive over time, if cooperation (and observation of cooperation) occurs over time.

SOME PROPOSED EXPERIMENTAL TESTS OF REPUTATION AND FORGIVENESS

To investigate the issue of negative reputation and the possibility of forgiveness and reputation repair within groups, we can use our definitions of *reputation* and *reputation history*. We propose several experimental tests that vary how people are informed of past cooperative behavior and whether communication is allowed. The proposed experiments ask different questions about the processes. Experiments are appropriate as we seek to test some general principles about reputations; we are not seeking to describe how reputations are used in any particular natural setting (at this point). Experiments are especially valuable for the tests of principles because they enable control and random assignment, thus ruling out alternative interpretations for results (Webster & Sell, 2014). [ii]

We vary the kind of communication that is possible to investigate what might be effective for repairing reputations. In one set of conditions, we suggest that, without a connection of behavior to a specific actor, reputations cannot form. This follows from our definitions of reputation: (1) connection by observers must be made between the specific actor and the specific behavior, and (2) this connection needs to be communicated. Further, in conditions in which reputations cannot change, because there is no mechanism for change, and some of those reputations are negative, cooperation becomes problematic. Initial negative reputations will drive noncooperation.

Consequently, we predict that in conditions in which observers cannot connect actions to actors, reputations cannot be developed and uncertainty will lead to lower levels of cooperation.

In conditions in which reputation is revealed, but there is no opportunity to repair a negative reputation, monitoring by group members will be heightened. Initially, cooperation will be relatively high (higher than in conditions where only past behaviors are known) but will decay over time. This occurs because any deviation from high cooperation will lead to a spiral of noncooperation.

In another set of conditions, we suggest that certain kinds of apologies or public acknowledgements are parts of a reputation in the same domain, and if this is the case, it is an additional element in the reputation set. In this case, repentance and acceptance of responsibility for wrong-doing would be most effective when it belongs to the same domain as the behavior in question. Additionally, the more observers who see this public acknowledgement, the stronger its effect. This prediction is simply an implication from the definition of *reputation* and *reputation history*.

As mentioned, the role of reputation is critical in helping solve social dilemmas. In the

proposed experiments, we investigate how the revelation of negative and positive reputations affects group cooperation. There are many potential ways to test reputations, and we offer several and briefly discuss positive and negative aspects of the design.

Experiment I. (Classification of Participants through Study 1, Repeated Interactions, and Information about Participants in Study 2)

This proposed experiment does not require deception, and we consider this an important advantage. It is decidedly more complex than other experiments we suggest and because it does not involve deception, is expensive for participant payment and will require more attention to eliminating alternative interpretations related to individuals' personality and "types." An especially strong aspect of this experiment is that it enables the examination of all group members' behavior over time, including those with noncooperative past behavior.

In Experiment 1, we will vary revelation of past behavior (reputation but not connected to specific actors), revelation of the reputation for each particular group member, and revelation with the opportunity for the low cooperative member to send messages. Given the definitions we created, this is varying the presence of information about past behavior (but not reputation), the reputations involved (that is the connection of the acts of cooperation with a particular actors), and finally the power of apology and acceptance of responsibility.

Development of Reputations (Study 1)

To develop actual reputations, we will ask a relatively large number of participants to engage in public goods studies that mirror the payoffs in previous studies. In these studies, participants will interact on an internet connection in which they will play in groups of 4 for around 7 trials. (The exact number of trials is not told to participants as this can create end effects.) There will be no feedback about others' choices in these settings; however, everyone

will receive their final earnings. We will divide up the distribution of those who participated based on the individual cooperation means and use the top 20%, the bottom 20%, and the middle 20%. We will take the top 20% of those participants who were cooperative, 20% who were most noncooperative, and middle 20%. We will then contact these participants and schedule them for Study 2: revelation of past behavior or reputation.

The design is a 2X3 factorial design. Power calculations based upon the initial pre-test experiments suggested that 22 groups per condition can detect medium effects.

The first independent variable is Reputation: (a) *Past behavior, No Revelation*, (b) *Reputation Revelation, and* (c) *Reputation Revelation, Communication Possible.* The second independent variable is Group Composition: (a) *1 noncooperator and 3 cooperators* and (b) *1 noncooperator and 3 middle cooperators.*

Independent Variables for Study 2 in Experiment I.

For the Group Composition variable, we will schedule those who have previously played public goods games. Based on their prior cooperative behavior, we will form groups with different compositions. In half the groups, there will be three people who are classified in the middle 20% of cooperators, and one person classified in the bottom 20% of cooperators. In the other half of the groups, there will be three people who are classified in the top 20% of cooperators, and one person classified in the bottom 20% of cooperators. The use of people who were either high cooperators or middle cooperators allows us to determine if acts of forgiveness are driven differently for different cooperative histories. For the Reputation variable, there are three variations. In the *Past Behavior, No Revelation* condition, only the fact that the group contains a certain number of cooperators, (e.g., medium level cooperators or noncooperators) will be revealed. No individual participant is identified with their past behavior, thus, by

definition there is no reputation, only information about past behavior. This condition allows us to estimate how information about composition vs information about specific individuals' reputations affects behavior. In the *Reputation Revelation* condition, individual participants are identified with their previous behavior. For example, while making their decisions, participants will see the classification of themselves and the other participants (e.g. Participant #1-Low Cooperator Previously or Participant #2 -High Cooperator Previously). In the *Reputation Revelation, Communication Possible* condition, Low Contributors will be allowed to send announcements to the other members at the beginning of the study. There 3 possible announcements that can be sent: (a) "I intend to contribute to the group fund", (b) "I realize my decisions in the previous study cost others in my group. I apologize, and I intend to contribute to the group fund", and (c) no announcement. Everyone will know that the noncooperator could send any of these three messages.[iii]

Dependent Variables

The dependent variables will include contributions to the group fund as a measure of cooperation. Measures of cooperation will be assessed on both the group and individual level. Additionally, we will include measures of feelings of closeness and identification with the group. We also intend to include questions about strategy and some open-ended questions about perceptions of others.

Predictions

We suggest that past behaviors are extremely difficult to change, most especially if that reputation is the only information known and if that information is not suspect. In the cases we consider here, the information about past behavior is accurate and everybody in the group knows that it is accurate. When the group composition but not the actual participant is revealed, this

creates uncertainty and considerable caution within the group. Our predictions are that cooperators, middle cooperators, and noncooperators will all begin contributions at a relatively high rate, but they will not be sustained. In essence, these groups will look like most other groups in such public goods settings in which cooperation begins at a high rate but steadily declines. In the *Reputation Revelation* condition, monitoring is at a heightened state. Those with negative reputations will attempt, at least initially, to recover their reputations through acts of cooperation, but suspicion will drive cooperation toward lower, more cautious levels. Cooperation in these conditions will start high as in the other conditions and be higher, in general, than the non-revelation conditions due to the careful monitoring by reputation. In the *Reputation Revelation-Communication Possible* condition, there will be great variance. Knowing that the noncooperators could make an announcement will change the group dynamics: both noncooperators and cooperators realize this. From the viewpoint of traditional game theory, any announcement is considered cheap talk, however monitoring is present, that is, everyone will see whether the announcement is used. No announcement will be read as a denial of responsibility and if the noncooperator chooses this, cooperation for the group will be lower than any other group. If the noncooperator chooses "I intend to contribute to the group fund", the groups will perform very similarly to groups that can monitor each other and know the reputations of each participant (*Reputation Revelation*). But, if the noncooperator chooses "I realize my decisions in the previous study cost others in my group and I apologize; I intend to contribute to the group fund", this will be read as a sign of acceptance of responsibility. In these cases, contributions will begin at high levels and stay at high levels. These groups will be the most cooperative (regardless of composition). This will occur because others in the group will put higher probabilities of cooperation on the previously noncooperative participants *because*

they have declared a negative (moral) judgment about themselves and because it carries a further commitment. Further this will enable other group members to demonstrate their ability to forgive or enable reputational repair. The literature on forgiveness suggests that the acceptance of responsibility is one of the most important components for the activation of forgiveness. Both the reputations of those with previous negative reputations and those with either mixed or positive reputations can be changed further toward cooperation. Because of this, these groups will contribute at the highest of levels of any of the groups. We also predict that initial reputation matters in all groups EXCEPT those in which the noncooperator assumes responsibility. That is, previously cooperative group members will contribute at higher levels than either the mid-level contributors or the noncontributors.

The reactions of the mid-level cooperators and high-level cooperators will also be explored. Since the reputations are based on public goods settings, it would be generally expected that participants would continue to act as these reputations place them; that is, high cooperators would try to sustain their high cooperative reputation while mid-level cooperators would not feel that same commitment, and so might be more sensitive to potential defections. These behaviors will translate to the feelings about the group itself with mid-level cooperators feeling less attached to the group than high level cooperators in general. However, in groups in which the low contributor accepts personal responsibility, all group members will feel more positive about the group, and express higher levels of group identity in the group. [iv]

Experiment II. A Test of Forgiveness within Groups

(Classification of Participants in Study 1, Study 2: Repeated Interactions but with no feedback about contributions)

This experiment has the same general format of Experiment I. Similar to Experiment I., it

utilizes a Study 1 public goods game to create the past behavior upon which reputations are developed. For the second study, groups are composed either of three high cooperators or three middle level cooperators. It utilizes deception to create the group member with the negative reputation.[v] The advantage of this experiment is that it ensures that we will have many negative reputation people in each category and would be less expensive to run. The disadvantages are that we have used deception, we do not have behaviors conditioned by feedback of others, and as a consequence of the design, we have no information about the participant with the noncooperative reputation. Consequently, this 2X3 factorial tests only forgiveness granted and not on the actions of the participant with the negative reputation.

Similar to our predictions for Experiment I, we predict that the highest level of contributions by group members will be in the case where the (fake) negative reputation participant apologizes and accepts responsibility. We predict that the lowest contributions will occur when the (fake) negative reputation participant has the ability to make an announcement but doesn't communicate at all.

As in Experiment I, we also intend to include questions about strategy and some open-ended questions about perceptions of others.

Experiment III. A Test of Forgiveness for Individuals

Experiment III is a vignette experiment that asks people to consider how they would behave in repeated game play. This experiment focuses upon the degree to which the participants would forgive the group member with the negative reputation. This experiment contains no study I. Participants will be randomly assigned and given information about the public goods game and will complete some practice trials. Then, they will be asked to contribute over a series of trials, without feedback of others. This vignette study is a study of how the

information about other group members changes the contribution of one group member, the participant. It asks a question only about information and its effect on one group member. The design would be a 5-condition experiment. The five conditions would be 1) *No Information about group composition (Baseline),* 2) *Knowledge that group is composed of one participant who demonstrated noncooperative behavior in the past, 3) Knowledge of past noncooperative behavior by one participant who chooses not to make an announcement, 4) Announcement of intension to contribute made by participant with a negative reputation, 5) Apology and acceptance of responsibility made by participant with a negative reputation.*

The experiment would be relatively easy to conduct as it could be all online. Participants would not be categorized through an initial study, but the baseline with no information allows us to gauge the effect of the different conditions. There is no deception. This is a 5-condition study in which the information about the participant with the negative reputation is varied: no information (as baseline); Information that one participant has been noncooperative in the past; Information that the noncooperative participant made no announcement about intention to contribute; information that the noncooperative participant announced they intended to contribute; information that the noncooperative participant apologized and accepted responsibility. Payment would be decided by taking a random decision made on a trial by the participant and pairing it with an actual group that had the same composition told to the participant. However, even following best practices for vignette studies (Aguinis and Bradley 2014), the problem is that often people do not really know how they might act, especially in settings that would be unfamiliar (Nisbett and Wilson 1977).

CONCLUSION

While the concept of reputation is important across different fields, the many different

interpretations of its meaning have hindered interdisciplinary cumulation. We offer some definitions which highlight how observers and their communication is important. This differentiates dyadic settings in which actions might be viewed as idiosyncratic from group settings in which observers communicate with others. Reputations develop based on past behavior AND communication of that past behavior to others. They are dynamic. Gossip and media are ways in which reputations are formed and modified, but so too are electronic evaluations, commonly used for businesses such as ratings for restaurants or for transactions. The observers who create the reputations might vary wildly in their methods to communicate and assess reputation, and these methods and how they are evaluated by others would be an important and vibrant area of investigation given misinformation, falsification, error, and lying.

But, there is no doubt that reputations are built on the past and affect the present and future in self-fulfilling ways. Those who have been judged as non-cooperative in the past are judged as non-cooperative in the future and therefore undesirable. This is especially problematic for groups in social dilemma settings. If those with histories of non-cooperation can never be trusted again, the only kind of options are ostracism, or group failure. Such an outcome is undesirable for many reasons: greater group divisions are created and monitoring such divisions to ensure that groups do not harm each other is also costly.

Questions of the restoration of reputation are critical. Reconciliation is necessary, as evolutionary theorists point out, for the survival of groups. In tremendous societal upheaval, we know reconciliation, even in the face of tremendous conflicts, has sometimes healed and allowed groups to work together. But it has also failed. So, it is critical to study and understand how reputations can be established and changed and how forgiveness might be granted.

REFERENCES

Abeler, J., Calaki, J., Andree, K., & Basek, C. (2010). The power of apology. *Economic Letters*, *107*(2), 233-235.

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17(4),* 351-371.

Amstutz, M. R. (2005). *The Healing of Nations: The Promise and Limits of Political Forgiveness.* Lanthan, MD: Rowman & Littlefield.

Anderson, C., & Shirako, A. (2008). Are individuals' reputations related to their history of behavior? *Journal of Personality and Social Psychology*, *94*(2), 320–333.

Aumann, R. J. (1987). Game theory. In: J. Eatwell, M. Milgate & P. Newman (Eds), *The new Palgrave: A dictionary of economics* (pp. 460–482). London: Macmillan Press.

Axelrod, R. (1980a). Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution, 24*(1), 3-25.

Axelrod, R. (1980b). More effective choice in the prisoner's dilemma. *Journal of Conflict Resolution. 24*(3), 379-403.

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons". *Evolution and Human Behavior, 25*(4), 209-220.

Barclay, P. & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings: Biological Sciences*, *274*(1610), 749-753.

Berger, J., Cohen, B. P., & Zelditch, Jr., M. (1972). Status characteristics and social interaction. *American Sociological Review, 37*(3), 241-255.

Binmore, K. (1992). *Fun and Games: A Text on Game Theory.* Lexington, MA: D.C. Heath and

Co.

Bohnet, I., & Huck, S. (2004). Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American Economic Review, 94*(2), 362–366.

Brewer, M. (2007). The importance of being we: Human nature and intergroup relations. *American Psychologist, November*, 728-738.

Cağin Bektaş, M. (2018). The Importance of social responsibility projects in reputation management: The role of public relations. *Global Media Journal*, *8*(16), 229–239.

Chartered Institute of Public Relations. (2019). What is PR? Retrieved from https://www.cipr.co.uk/content/policy/careers-advice/what-pr

Clare, J. & Danilovic, V. (2011). Multiple audiences and reputation building in international conflicts. *Journal of Conflict Resolution, 54*(6), 860-882.

Darby, B. W., & Schlenker, B. R. (1989). Children's reactions to transgressions: Effects of the actor's apology, reputation and remorse. *British Journal of Social Psychology*, *28*(4), 353–364.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169–193.

Deutsch, M. & Gerard, H. B. (1955). A study of normative and informational influences upon individual judgment. *Journal of Abnormal and Social Psychology, 51*(3), 629-636.

de Waal, F. (1996). Good natured: the origins of right and wrong in humans and other animals. Cambridge, MA: Harvard University Press.

de Waal, F. & Pokorny, J.J. (2005). Primate conflict and its relation to human forgiveness. In E. L. Worthington, Jr. (Ed.), *Handbook of Forgiveness* (pp. 17-40). London: Routledge.

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology, 8*(2), 100-110.

Denko, M. K., Sun, T., & Woungang, I. (2011). Trust management in ubiquitous computing: A Bayesian approach. *Computer Communications*, *34*(3), 398–406.

Emler, N. (1990). A social psychology of reputation. *European Review of Social Psychology*, *1*(1), 171–193.

Emler, N., St. James, A., & Faucheux, A. (2013). Reputation as a social instrument. *Communications*, *2*(93), 85–99.

Eriksson, K. & Strimling, P. (2012). The hard problem of cooperation. *PLOS ONE*, *7*(7), e40325.

Fehr, E. & Gäechter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review, 90*(4), 980–994.

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.

Fehr, R. & Gelfand, M. J., Nag, M. (2010). The road to forgiveness: A meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin, 136*(5), 894–914.

Feinberg, M., Willer, R., & Shultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science, 25*(3), 656-664.

Fine, G. A. (1996). Reputational entrepreneurs and the memory of incompetence: Melting supporters, partisan warriors, and images of President Harding. *American Journal of Sociology*, *101*(5), 1159-1193.

Fine, G. A. (2019). Moral cultures, reputation work, and the politics of scandal. *Annual Review of Sociology*, *45*(1), 247-264.

Fombrun, C., & Shanley, M. (1990). What's in a name: Reputation building and corporate strategy. *Academy of Management Journal*, *33*(2), 233-258.

Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica, 54*(3), 533–554.

Fullerton, J., & Kendrick, A. (2017). Country reputation as a moderator of tourism advertising effectiveness. *Journal of Marketing Communications*, *23*(3), 260–272.

Gibler, D. M. (2008). The costs of reneging: Reputation and alliance formation. *Journal of Conflict Resolution, 52*(3), 426-454.

Gibson, D., Gonzales, J. L., & Castanon, J. (2006). The Importance of reputation and the role of public relations. *Public Relations Quarterly, 51*(3), 15–18.

Govier, T. (2006). *Taking Wrongs Seriously: Acknowledgment, Reconciliation and the Politics of Sustainable Peace*. Amherst: NY: Humanity Books.

Griskevicius, V., Tybur, J. M., & Van den Bergh, B. (2010). Going green to be seen: Status, reputation, and conspicuous conservation. *Journal of Personality and Social Psychology, 98*(3), 392–404.

Guisinger, A. & Smith, A. (2002). Honest threats: The interaction of reputation and political institutions in international crises. *Journal of Conflict Resolution, 46*(2), 175-200.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(May), 998-1002.

Hayes, T. L., Hogan, R., & Emler, N. (2016). The Psychology of Character, Reputation, and Gossip. In I. Fileva (Ed.), *Questions of Character*. Oxford University Press.

Hayner, P. B. (2010). *Unspeakable Truths: Transitional Justice and the Challenge of Truth Commissions.* New York, NY: Routledge.

Harsanyi, J. C. (1967). Games with incomplete information played by 'Bayesian' players, I–III.

Part I. The Basic Model. *Management Science*, *14*(3), 159–182.

Harsanyi, J. C. (1968a). Games with incomplete information played by 'Bayesian' players, I–III. Part II. Bayesian Equilibrium Points. *Management Science*, *14*(5), 320–334, 486–502.

Harsanyi, J. C. (1968b). Games with incomplete information played by 'Bayesian' players, I–III. Part III. The Basic Probability Distribution of the Game. *Management Science*, *14*(7), 486–502.

Halbwachs, M. (1980). *The Collective Memory*. New York, NY: Harper Colophon.

Ho, B. (2012). Apologies as signals: with evidence from the trust game. *Management Science*, *58*(1), 141-158.

Hughes, P. M. & Warmke, B. (2017). Forgiveness. In E. N. Zalta (ed.)*The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition) https://plato.stanford.edu/archives/sum2017/entries/forgiveness/).

Jazaieri, H., Logli Allison, M., Campos, B., Young, R. C., & Keltner, D. (2018). Content, structure, and dynamics of personal reputation: The role of trust and status potential within social networks. *Group Processes and Intergroup Relations*, *0*(0), 1–20.

Jervis, R. (1988). War and Misperception. *The Journal of Interdisciplinary History*, *18*(4), 675-700.

King, B. G., & Whetten, D. A. (2008). Rethinking the relationship between reputation and legitimacy: A social actor conceptualization. *Corporate Reputation Review*, *11*(3), 192–207.

Klabi, H., Mellouli, S., & Rekik, M. (2014). A reputation based electronic government procurement model. *Government Information Quarterly*, *35*(4), 543-553.

Kollock, P. (1994). The emergence of exchange structures: An experimental study of

uncertainty, commitment, and trust. *American Journal of Sociology, 100*(2), 313-345.

Kollock, P. (1998). Social Dilemmas: The anatomy of cooperation. *Annual Review of Sociology, 24*, 183-214.

Kottasz, R., & Bennett, R. (2016). Managing the reputation of the banking industry after the global financial crisis: Implications of public anger, processing depth and retroactive memory interference for public recall of events. *Journal of Marketing Communications, 22*(3), 284–306.

Kuwabara, K. (2015). Do reputation systems undermine trust?: Divergent effects of enforcement type on generalized trust and trustworthiness. *American Journal of Sociology, 120*(5), 1390-1428.

Lang, G. E. & Lang, K. (1988). Recognition and renown: The survival of artistic reputation. *American Journal of Sociology, 94*(1), 79-109.

Lange, D., Lee, P., & Dai Y. (2011). Organizational reputation: A review. *Journal of Management, 37*(1), 153-184.

Lawler, E. J., Thye, S. R., & Yoon, J. (2000). Exchange and group cohesion in productive exchange. *American Journal of Sociology, 106*(3), 616-57.

Lawler, E. J., Thye, S. R., & Yoon, J. (2009). *Social commitments in a depersonalized world*. New York, NY: Russell Sage Foundation, Inc.

Lewis, B. (1975). *History: Remembered, Recaptured, Invented*. Princeton, New Jersey: Princeton University Press.

McCullough, M. E., & Hoyt, W. T. (2002). Transgression-Related motivational dispositions: Personality substrates of forgiveness and their links to the big five. *Personality and Social Psychology Bulletin, 28*(11), 1556–1573.

McCullough, M. E., Rachal, K. C., Sandage, S. J., Worthington, Jr., E. L., Brown, S. W., & Hight, T. L. (1998). Interpersonal forgiving in close relationships: II. Theoretical Elaboration and Measurement. *Journal of Personality and Social Psychology*, *75*(6), 1586–1603.

McDonnell, M.-H. & King, B. (2013). Keeping up Appearances: Reputational Threat and Impression Management after Social Movement Boycotts. *Administrative Science Quarterly, 58*(3), 387–419.

McDonnell, M.-H. & King, B. G. (2018). Order in the court: How firm status and reputation shape the outcomes of employment discrimination suits. *American Sociological Review*, *83*(1), 61-87.

Molm, L. D., Takahashi, N., & Peterson, G. (2000). Power in negotiated and reciprocal exchange. *American Sociological Review*, *64*(6), 876-890.

Nisbett, R. E., Wilson T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-258.

Nowak, M. A., Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 393*, 573–577.

Nowak, M. A., Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*, 1291–1298.

Ohtsubo, Y. & Yagi, A. (2015). Relationship value promotes costly apology-making: testing the valuable relationships hypothesis from the perpetrators perspectives. *Evolution and Human Behavior*, *36*(3), 232-239.

Olson, M., Jr. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Ostrom, E. (1990). *Governing the Commons*. New York, NY: Oxford.

Ostrom, E. (1998). A Behavioral Approach to the Rational Choice Theory of Collective Action, Presidential Address, American Political Science Association, 1997. *American Political Science Review, 92*(1), 1-22.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2010). Evolutionary Psychology and Criminal Justice: A Recalibrational Theory of Punishment and Reconciliation. In H. Høgh-Olesen (Ed.), *Human Morality and Sociality: Evolutionary & Comparative Perspectives* (pp. 72-131). Palgrave Macmillan.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior, 33*(6), 682-695.

Radzik, L. (2010). Moral bystanders and the virtue of forgiveness. In C. R. Allers & M. Smit (Ed.), *Forgiveness in Perspective* (pp. 66-69). Amsterdam: Rodopi.

Raub, W. & Weesie, J. (1990). Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology, 96*(3), 626-654.

Roberts, P. W., & Dowling, G. R. (2002). Corporate reputation and sustained superior financial performance. *Strategic Management Journal, 23*, 1077-1093

Rogosch, F. A., & Newcomb, A. F. (1989). Children's perceptions of peer reputations and their social reputations among peers. *Child Development, 60*(3), 597–610.

Sallot, L. M. (1993). *The Effects of Motive, Communication Style, and Licensing on the Reputation of Public Relations: An Impression Management Perspective* (Unpublished doctoral dissertation). University of Florida, Florida.

Sartori, A. (2005). *Deterrence by Diplomacy*. Princeton, NJ: Princeton University Press.

Schmitt, R. L. & Leonard II, W. M. (1986). Immortalizing the self through sport. *American Journal of Sociology, 91*(5), 1088-1111.

Sechser, T. S. (2010). Goliath's curse: Coercive threats and asymmetric power. *International Organization, 64*(Fall 2010), 627-660.

Sechser, T. S. (2018). Reputations and signaling in coercive bargaining. *Journal of Conflict Resolution, 62*(2), 318-345.

Sell, J. (1988). Types of public goods and free-riding. In: E. J. Lawler & B. Markovsky (Ed.), *Advances in Group Processes* (Vol. 5, pp. 119–140). Greenwich, CT: JAI Press.

Sell, J. (1997). Gender, strategies, and contributions to public goods. *Social Psychology Quarterly*, 60(3), 252-265.

Sell, J. (2018). Definitions and the development of theory in social psychology. *Social Psychology Quarterly, 81* (1), 8-22.

Simpson, B. & Willer, R. (2015). Beyond altruism: Sociological foundations of cooperation. *Annual Review of Sociology, 41*(1), 43-63.

Simpson, B., Willer, R., & Harrell, A. (2017). The enforcement of moral boundaries promotes cooperation and prosocial behavior in groups. *Scientific Reports, 7,* 42844.

Snyder, J. (1984). Civil-Military Relations and the Cult of the Offensive, 1914 and 1984. *International Security, 9*(1), 108-146.

Sorenson, O., (2014). Status and reputation: Synonyms or separate concepts? *Strategic Organization, 12*(1), 62-69.

Stets, J. & Carter, M. J. (2012). A theory of the self for the sociology of morality. *American Sociological Review, 77*(1), 120-140.

Sugden, R. (1986). *The Economics of Rights, Co-operation and Welfare*. Oxford, UK: Blackwell

Publishers Ltd.

Suzuki, S. & Akiyama, E. (2005). Reputation and the evolution of cooperation in sizable groups. *Proceedings. Biological sciences*, *272*(1570), 1373–1377.

Tafreschi, O., Mähler, D., Fengel, J., Rebstock, M., & Eckert, C. (2008). A reputation system for electronic negotiations. *Computer Standards & Interfaces*, *30*(6), 351–360.

Thye, S. (2000). A status value theory of power in exchange relations. *American Sociological Review, 65*(3), 407-432.

Tingley, D. H. & Walter, B. F. (2011a). The effect of repeated play on reputation building: An experimental approach. *International Organization, 65*(2), 343-365.

Tingley, D. H. & Walter, B. F. (2011b). Can cheap talk deter?: An experimental analysis. *Journal of Conflict Resolution, 55*(6), 996-1020.

Wade, N. G., Hoyt, W. T., Kidwell, J. E. M. & Worthington, E. L. Jr. (2014). Efficacy of psychotherapeutic interventions to promote forgiveness: A Meta-analysis. *Journal of Consulting and Clinical Psychology, 82*(1), 154-170.

Wade, N. G., & Worthington, E. L., Jr. (2003). Overcoming unforgiveness: Is forgiveness the only way to deal with unforgiveness? *Journal of Counseling & Development, 81,* 343–353.

Wang, Y., & Vassileva, J. (2003). Bayesian network trust model in peer-to-peer networks. In G. Moro, C. Sartori, and M.P. Singh (Eds.) *Agents and Peer-to-Peer Computing: Second International Workshop*, *AP2PC 2003* (pp. 23–34). Berlin, Germany: Springer-Verlag.

Webster, M. & Sell, J. (2014). Why do experiments? In M. Webster & J. Sell (Eds.), *Laboratory Experiments in the Social Sciences* (pp. 5-22). New York, NY: Elsevier.

Wiegand, K. E. (2011). Militarized territorial disputes: States' attempts to transfer reputation for

resolve. *Journal of Peace Research, 48(*1), 101-113.

Weisiger, A. & Yarhi-Milo, K. (2015). Revisiting reputation: How past actions matter in international politics. *International Organization, 69*(2), 473-495.

Willer, R. (2009). Groups reward individual sacrifice: the status solution to the collective action problem. *American Sociological Review*, *74*(1), 23-43.

Williamson, O. E. (1981). The economics of organization: The transaction cost approach. *American Journal of Sociology*, *87*(3), 549-77.

Williamson, I., Gonzales, M. H., Fernandez, S., & Williams, A. (2014). Forgiveness aversion: Developing a motivational state measure of perceived forgiveness risks. *Motivation and Emotion*, *38*(3), 378–400.

Wilson, R. (1985). Reputations in Games and Markets. In A. Roth (Ed.) *Game Theoretic Models of Bargaining* (pp. 27-62). Cambridge: Cambridge University Press.

Wilson, R. K., & Sell, J. (1997). Cheap talk and reputation in repeated public goods settings. *Journal of Conflict Resolution, 41*(5), 695-717.

Worthington, E. L., Jr. (Ed.). (2005). Handbook of forgiveness. New York, NY: Brunner/Routledge.

Wu, C. X. & Wolford, S. (2018). Leaders, states, and reputations. *Journal of Conflict Resolution, 62*(10), 2087-2117.

Xiong, L., & Lui, L. (2004). Peer trust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Transactions on Knowledge and Data Engineering, 16*, 843–857.

[i] The fact that cooperative actors are rewarded brings up the issue of status and how status might be different from reputation. McDonnell and King (2018) differentiate status from reputation as do Simpson and Willer (2015). Both sets of researchers reference expectation states formulations in reference to status and define it

as an attribute related to position and so operates through a hierarchy (Berger, Cohen, & Zelditch, 1972; Thye, 2000). For example, a person might occupy a high status position in a company, and thereby have influence, but that same person might have varied reputations across different domains: She might be a cooperative person within the company based on past behavior, but she might also be ruthless with competitors. So, while status is relative to the position, such that a person might be higher status in one setting and lower status in another, it does not necessarily evoke expectations based on past behavior and domain specificity the way reputation does. Additionally, reputation is not necessarily based upon a hierarchy, as is the case with status.

[ii] We know that, under some conditions, characteristics of the participants can create differences in decision-making behavior (see for example, Sell, 1997). Because this is the case, we intend to conduct all the studies either online or in settings to avoid face-to-face decision making.

[iii] We do not know how previous noncooperators will choose and, in fact, we will explore their reasoning and strategy at the end of this study. We anticipate that we may need up to 60 groups in this condition.

[iv] Another potential calibration study would ascertain how important participants perceive their reputations to be. All participants will have been in Study I, the public goods game online with no feedback. Participants will be contacted and given information about their past behavior (i.e., their reputation.). They will be told that they will be interacting in an experiment in which their reputation is revealed. However, they will be further informed that they can change from this scheduled study to a study in which their reputations would not be revealed. As a caveat, they are also told that because we will have to change scheduling which affects payment of researchers, there will be a slight change fee that can be subtracted from the tokens from their future earnings. They will then be asked if they would like to change and if so, if how much (how many tokens) they are willing to pay for the change. The prediction is that noncontributors from Study 1 will choose not to be in a Reputation Revealed study II., and willing to pay for it. It may be that High Cooperators and Middle Cooperators will also choose to change, but they will do so at a much lower rate.

[v] It is possible to run this study and technically not be considered "deceptive" if we add in our consent form a statement to the effect that it is possible that some of the information provided may be misleading. While this is technically not deception, it is problematic from a number of viewpoints. For example, it probably creates enough doubt for participants that could affect what information they use or don't use.