I'm sorry if you are: The risk of apologizing first

Shereen J. Chaudhry*
University of Chicago Booth School of Business
5807 S. Woodlawn Ave
Chicago, IL 60637
chauds@chicagobooth.edu

Valeria Burdea
Ludwig-Maximilian University Munich
Seminar for Economic Theory
Ludwigstraße 28 / 04a (Rgb.)
80539 Munich
valeria.burdea@econ.lmu.de

*Corresponding author

May 2022

Abstract

Apologies are powerful tools for reconciliation, so understanding (and addressing) barriers to apologizing can enhance long-term cooperation. Existing research on apologies leaves critical blind-spots by focusing on one-sided blame conflicts: We find that most unresolved conflicts involve *mutual* blame. Using a game theoretic framework, we highlight strategic considerations unique to mutual blame conflicts, identifying an unexplored apologizing barrier: the risk of not receiving a return apology. We propose that people expect the pain of apologizing *first* to be lessened when followed by a return apology, even more so than when followed by forgiveness. We confirm these predictions across seven pre-registered studies that include a game with monetary incentives, event recall paradigms, and scenarios ranging from workplace conflicts to international diplomatic ones. This work illustrates the existence of strategic concerns in apologizing that are exclusive to the ubiquitous mutual blame conflicts, and highlights an unexplored avenue for research in conflict management.

Keywords: apology, blame, conflict, reconciliation, relationships, social communication, game theory

1. Introduction

Maintaining harmony and cooperation in ongoing relationships with other people requires knowing how to move past the inevitable offenses, violations of trust, and conflicts of interest without the deterioration of the relationship—that is, knowing how to reconcile. But consider how often people fail at this, leading to significant consequence in their relationships. Why is it not uncommon to hear about close friendships that dissolve after a seemingly mundane squabble? Or family members who have treated each other to decades of radio silence? What about the long-standing colleagues who seem to avoid each other at every turn for years on end? In asking people to recall such unresolved conflicts from their lives (Studies 1c & 4), we noticed a remarkable pattern: The majority of these disputes are characterized by mutual blame—that is, both parties bear some amount of blame (93% in Study 1c, N = 61; 63% in Study 4, N = 283). Despite knowing they are also to blame, most mutual blame combatants report that neither person has apologized (58% of mutual blame conflicts in Study 4). That is, an apology stalemate follows most mutual-blame conflicts.

This is a critical observation because prior scholarship has established that the act of apologizing is a powerful tool for reconciliation (Darby & Schlenker, 1982, 1989; Fehr & Gelfand, 2010; Kirchhoff, Wagner, & Strack, 2012; Lewicki, Polin, & Lount, 2016; Scher & Darley, 1997; Schumann, 2018; Tomlinson, Dineen, & Lewicki, 2004). What is keeping people from apologizing in the conflicts described above? Could there be features unique to mutual blame conflicts that create barriers to apologizing? Because existing research has focused exclusively on one-sided blame conflicts (see Lewis, Parra, & Cohen, 2015, for a review), we argue that these questions highlight critical blind-spots in the understanding of apologies, and thus, of reconciliation solutions.

To illustrate, consider the following scenario about an interpersonal infraction: After a work conference, Fran and Saul—colleagues who interact frequently at work—each have tickets on the same flight back to their home city, so they decide to carpool to the airport in Fran's rental car. Saul gets in Fran's car 20 minutes late because he waited until the last minute to pack his bag. During the drive, Fran refuses to use GPS to find the airport resulting in a 20-minute detour that causes them to miss their flight.

If the delay was only 20 minutes in total, they could have caught the flight, but the *combination* of their behaviors, leading to a 40-minute delay, causes Fran and Saul to miss their flight—that is, the blame is *mutual*. While both are furious, each person has something they could apologize for. Due to the sequential nature of conversation, one person must go first. Would it be easier for Fran to apologize first or second? If Fran is the first to apologize, would she be equally happy in the case that Saul simply forgave her ("It's okay") as in the case that he reciprocated the apology ("I'm sorry, too")? Or would the former be unsatisfying, leaving her to harbor negative feelings toward Saul? Would knowing Saul would reciprocate the apology make Fran more likely to take the first step?

The present research addresses the open questions posed above, filling in the gaps in the literature, by applying a game-theoretic framework to the decision to apologize in mutual blame conflicts, which we describe in the next section. In doing so, we identify strategic considerations and risks that are unique to mutual blame conflicts and that represent unexplored barriers to apologizing. In particular, we hypothesize that apologizing first and getting a return apology is perceived to feel better than apologizing alone, making apologizing first a risky choice. Moreover, a return apology has a unique role in these instances such that alternative response communications, like forgiveness, cannot substitute its perceived benefits. These differences in perceived costs imply that people may display "conditional initiation" around apologizing, wherein they are more likely to apologize first if the perceived likelihood of getting a return apology is higher. These perceived costs also imply that people will be more willing to apologize second because, in that case, the risk of being the only one to apologize is eliminated.

We test these claims across seven pre-registered studies that utilize multiple methods including real behavioral interactions with monetary incentives, event recall paradigms, and scenarios ranging from interpersonal work conflicts to international diplomatic ones. Studies 1a-c confirm across three different populations that on average people expect to feel positive when they apologize and get a return apology but negative when they apologize alone. We also show that receiving forgiveness after an initial apology is expected to feel worse than receiving a return apology. Studies 2-4 document conditional initiation across three different domains, showing that the willingness to apologize first is persistently sensitive to

the perceived likelihood of getting a return apology. Studies 4 and 5 demonstrate that people are more likely to apologize second than first, even after controlling for blame perceptions.

By addressing unanswered questions around apologizing in mutual blame conflicts, the present work expands our understanding of the dynamics of reconciliation as well as when people will and will not be willing to apologize. The strategic concerns we highlight suggest that mutual blame conflicts may be more prone to remaining unresolved than unilateral blame conflicts—if both parties are too afraid to go first, an apology stalemate will result, even if both would prefer to apologize and reconcile. However, the framework also implies clear solutions for improving reconciliation (e.g., correcting miscalibrated beliefs about the likelihood of a return apology).

2. Theoretical framework and hypotheses

Returning to the mutual blame example, imagine that Fran is considering apologizing. We will refer to Fran as the first mover (notice her name starts with "F" for "first") and Saul as the second mover (notice his name starts with "S" for "second"). We can model Fran's decision space using a two-stage sequential-move game in which each player has two actions (apologize or don't apologize), but where players only know their own expected payoffs (i.e., how they expect they will feel for a given outcome). Figure 1 displays Fran's game tree. If Fran does not apologize, there are two possibilities (not shown): She can end up in the case where neither she nor Saul apologize, and no costs or benefits from apologizing are incurred. We call this an "apology stalemate." Fran could also end up in the situation where Saul apologizes first, giving her the flexibility to reciprocate the apology only if she wants to (i.e., the benefits outweigh the costs). This means that choosing not to apologize can only have an upside (relative to the status quo) for Fran, rendering this action riskless. Thus, we can reasonably argue that Z, the expected value of not apologizing (the right branch of Fran's tree in Figure 1), is equivalent to the status quo from which she is making her decision.

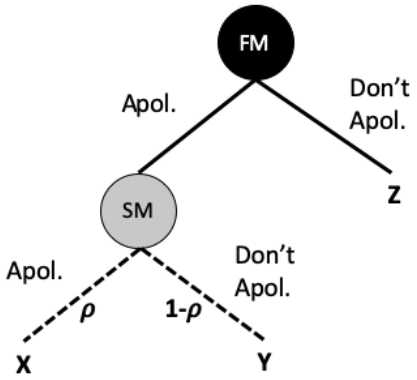


Figure 1. Game tree faced by Fran, the first mover (FM; black node), who does not know Saul's, the second mover's (SM; gray node), payoffs. If FM apologizes, her expected payoff is X, if SM apologizes, and Y, if SM does not. If FM does not apologize, her expected payoff is Z, which represents the expected value across all alternative outcomes. The alternative includes the possibility that neither person apologizes as well as the possibility that SM apologizes first.

Alternatively, Fran can deviate from the status quo by apologizing (the left branch of Fran's tree in Figure 1), and this could either result in "apologizing together," if Saul reciprocates the apology, or "apologizing alone," if he does not. We argue that for Fran, apologizing constitutes a risky choice because of two reasons. The first one is that the expected payoffs from apologizing are not perceived as equal in the eyes of the decision maker—Fran anticipates that she would strictly prefer apologizing together to apologizing alone. Our first hypothesis follows.

> **H1:** The expected payoff from apologizing and getting a return apology is higher than the expected payoff from apologizing and not getting a return apology (i.e., X > Y).

This hypothesis rests on the earlier observation that an apology involves admitting wrongdoing, which can have negative consequences for the apologizer's reputation (Kim et al., 2013, 2006) and self-image (Chaudhry & Loewenstein, 2019; Howell et al., 2011; Okimoto et al., 2013; Schumann, 2014).

But, if Fran perceives the benefits of reconciliation or guilt-alleviation to be large enough to overcome these costs, why would she care if Saul were to return the apology?

One possibility is that in the absence of a return apology from Saul (leading to the outcome Y in Figure 1), there is plausible deniability about Saul's guilt (Chaudhry & Loewenstein, 2019; De Freitas, Thomas, DeScioli, & Pinker, 2019; Lee & Pinker, 2010; Pinker, Nowak, & Lee, 2008). Saul leaves room to deny responsibility or present a different interpretation of events from that of Fran. Having apologized, Fran does not have this flexibility, and may appear to be taking all the blame—more blame than she feels she deserves. Thus, a lone apology transforms the relationship between the two individuals to one of victim-and-transgressor, putting the power to restore trust, as well as the right to punish, in the hands of the counterpart who didn't apologize (Bies & Tripp, 1996, pp. 259-260). It also prevents the apologizer from having the right to retaliate, and may even incite third parties to punish the apologizer (Jordan, Hoffman, Bloom, & Rand, 2016). However, a return apology (leading to outcome X in Figure 1) would mean both are victim and transgressor, effectively splitting the blame and reducing the asymmetry in the relationship obligations between the two. With a return apology, Saul does not have the moral high ground over Fran to punish asymmetrically, but rather the best strategy might be mutual absolution.

An alternative reason Fran's perceived benefit of an apology is different when she apologizes alone versus together with Saul is that she views the return apology as a form of forgiveness. If that is the case, then the act of Saul apologizing in return could be equivalent to him accepting her apology and forgiving her. Previous research has shown that the willingness to apologize in one-sided blame conflicts is positively related to the perceived likelihood of being forgiven (Leunissen et al., 2012), suggesting that the cost of apologizing is reduced when receiving forgiveness, and perhaps similarly, when receiving a return apology. However, if our theorizing is correct, then simply receiving forgiveness (without a return apology) in response to an apology would not be a sufficient substitute for a return apology. If parties in mutual blame conflicts want to avoid the relationship transforming into an asymmetric victim-transgressor relationship, potential apologizers will view forgiveness as inferior to a return apology. We thus formulate the following hypothesis.

H2: The expected payoff from apologizing and getting a return apology is higher than the expected payoff from apologizing and being forgiven.

The second reason apologizing can be a risky choice is that the expected value of the status quo often falls in between the expected value of the two outcomes of apologizing (i.e., X > Z > Y). Like taking on a mixed gamble, Fran expects that apologizing can make her either better off or worse off than the status quo—the cost of apologizing will only be outweighed by a return apology from Saul. As described above, apologizing alone may make Fran look like the lone transgressor and transfer the moral high ground to Saul. Furthermore, reconciliation may still be out of reach if she feels the need for an apology from Saul. As a result, not apologizing at all may leave her better off than apologizing alone (i.e., Z > Y). In contrast, a return apology could mitigate the costs of apologizing alone by splitting the blame and leveling the moral playing field, as described above. Moreover, the experience of apologizing together could create benefits that make it more preferred than the status quo of not apologizing (i.e., X > Z). Given that two individuals in a conflict often do not have the same interpretation of what exactly transpired (Baumeister, Stillwell, & Wotman, 1990; Feeney & Hill, 2006; Mikula, Athenstaedt, Heschgl, & Heimgartner, 1998), a return apology would validate the first mover's belief that blame is mutual and help to create a sense of *shared reality* between the two, i.e., a sense that both people share the same interpretation and attitude about what transpired (Higgins & Pittman, 2008), and this is intrinsically enjoyable (Echterhoff, Higgins, & Levine, 2009; Rossignac-Milon, Bolger, Zee, Boothby, & Higgins, 2021). Moreover, by addressing both people's needs as victims, apologizing together might be more effective at achieving reconciliation than apologizing alone.

This preference structure (i.e., X > Z > Y) implies that the choice to apologize first is sensitive to the perceived likelihood (represented by the term ρ in Figure 1) of getting X, the expected value of the better outcome, i.e., a return apology. Our next hypothesis reflects this idea.

H3: People display "conditional initiation" in that they are more likely to apologize first the more likely they think it is that they will receive a return apology.

In other words, under this preference structure, there is no dominant strategy—Fran's optimal strategy will depend on what she thinks Saul's payoffs are, and therefore, what action she thinks Saul is more likely to take. If Fran believes Saul's payoffs are similar to those in a prisoner's dilemma, where Saul benefits more by choosing not to return an apology (e.g., because Saul believes he is blameless), the best option for Fran is to avoid apologizing first, even if that results in an apology stalemate. If she believes Saul's payoffs are similar to those in a stag hunt game, where Saul benefits most when he reciprocates an apology (e.g., because he believes he should own up to his part of the conflict), the best option for Fran is to apologize. In this setup, uncertainty about getting a return apology constitutes a barrier to apologizing first, so information about the second mover's payoffs (i.e., the second mover's likely action) is paramount in deciding whether to apologize. We call this behavior pattern "conditional initiation," and we hypothesize that enough people display this behavior such that manipulating the perceived likelihood of getting a return apology (ρ) will on average lead to a change in the willingness to apologize first. Thus, in identifying this barrier, we also highlight a *potential solution*: conveying positive information about the counterpart's willingness to return an apology.

If apologizing first is risky at least in part because of the possibility of ending up being the only one to apologize, then apologizing second represents a less risky choice—the risk of apologizing alone is eliminated for a person who has received an apology, i.e. $\rho = 1$. As a result, we make an additional prediction that we test in this paper:

H4: People are more willing to apologize second (i.e., after the other person already apologized) than first, even after controlling for perceived relative blame.

This framework also makes predictions that we do not test here about how order affects perceived sincerity of the apology as well as perceived blame distribution. We outline these in the General Discussion and offer some initial evidence.

We ran seven studies using a variety of contexts to test the four hypotheses delineated above. Table 1 summarizes our findings. This is followed by a detailed description and discussion of each study and its results. All studies were approved by a university Institutional Review Board, and informed consent was obtained at the beginning of all experiments. We pre-registered all seven studies on AsPredicted.org. We report all manipulations, measures, and exclusions. In cases when our analyses or exclusions slightly differ from the pre-registration, we explain why. All pre-registrations, original surveys, data, and code to replicate the main and supplementary results are publicly available and have been deposited in the Open Science Framework (https://osf.io/jbg8f/?view_only=b1d12f0f407b457882fcd949f7f59b5c). Additional methodological details and analyses can be found in the appendices in the accompanying online supplement.

Table 1. Overview of studies

Study	Type	N	Main Finding(s)	Average feelings/apology likelihood by condition [a,b]		Statistical Test	Effect Size
				Alone (Y) / Forgiven	Together (X)		
1a	Event recall (MBA & undergraduate students)	42	Participants expected "apologizing alone" to feel worse than "apologizing together" (H1).	-18.76	6.98	$t(41) = 11.51$, $p < .001$	$d = 1.78$
			Participants expected getting a return forgiveness message to feel worse than "apologizing together" (H2).	-12.60	6.98	$t(41) = 6.94$, $p < .001$	$d = 1.07$
1b	Event recall (international professionals)	156	Same as Study 1a (H1).	-2.11	5.08	$t(155) = 6.40$, $p < .001$	$d = 0.51$
			Same as Study 1a (H2).	2.24	5.08	$t(155) = 2.79$, $p = .006$	$d = 0.22$
1c	Event recall (general public)	56	Same as Study 1a (H1).	-19.68	10.25	$t(55) = 14.50$, $p < .001$	$d = 1.94$
				Low	High		
2	Real lab-based conflict	280	Participants were more likely to apologize first in a free form message when they perceived their partner to be more likely to apologize in return (HIGH) than when they perceived them to be less likely to do so (LOW; H3).	17.14%	25.71%	X^2 (1, N = 280) = 3.05, $p = .0805$	$\phi = 0.10$
3	Leadership scenario (US citizens)	600	Participants were more likely to support the US president's initiating an apology for the US's past behavior towards Japan in WWII when they thought it was more likely that Japan's prime minister would apologize in return for Japan's past WWII behavior (HIGH) than when they thought this was less likely (LOW; H3).	8.62	14.41	$t(598) = -3.71$, $p < .001$	$d = 0.30$
4	Event recall	150	Participants reported being less likely to apologize first when they perceived the other person would be unlikely (compared to likely) to return the apology (H3).	-23.87	-4.14	$t(148.00) = 9.53$, $p < .001$	$\beta = 1.08$
				First	Second		
			Participants were more likely to apologize second than first using a continuous response scale (H4)	-15.87	9.29	$t(149.00) = 15.19$, $p < .001$	$\beta = 1.09$
5	Workplace scenario	800	Participants were more likely to apologize second than first when selecting from six discrete message choices (H4).	39.75%	68.25%	$t(797) = 8.07$, $p < .001$	$\beta = 0.53$

[a] Values for studies 1a-1c represent average expected feelings on a continuous sliding scale from -30 to 30.
[b] Values for studies 2-5 represent either proportion of participants apologizing (%) or average likelihood to apologize on a continuous sliding scale from -30 to 30.

3. Study 1: Anticipated feelings from apologizing alone versus together

Study 1 was designed to test whether people anticipate they will feel better after apologizing and receiving a return apology (apologizing together) than after apologizing and not receiving a return apology (apologizing alone; H1). We also test whether people expect receiving a return apology to feel better than receiving forgiveness after apologizing first (H2). To this end, we conducted three event recall studies with three different populations (1a, 1b, and 1c) in which we asked participants to recall an unresolved mutual blame conflict from their life. We focused on unresolved conflicts because we wanted to identify situations where apologies were least likely to have already been exchanged—our main test involves asking participants to imagine, regardless of what has already happened, that neither person has yet apologized.

Additionally, we wanted to get a sense of the proportion of real, unresolved conflicts that involve mutual blame, as opposed to one-sided blame. Unresolved conflicts represent the situations most in need of reconciliatory processes. Understanding whether these mostly constitute mutual blame cases, and also what proportion of these cases are those in which neither person has yet apologized, could help highlight the potential value of our proposed framework. To this end, in Study 1c, we added an additional phase before the main procedure asking people to recall *any* unresolved conflict from their life. We examined what proportion of participants spontaneously thought of mutual blame conflicts. This study's hypotheses and design were pre-registered at https://aspredicted.org/OSF_IVV (Study 1a & 1b) and https://aspredicted.org/MKR_VQS (Study 1c).

Methods

Initial procedure for Study 1c only. To assess the share of unresolved conflicts that involve mutual blame, we first asked participants to think of any unresolved conflict from their life. Participants indicated who was to blame for the conflict [multiple choice: "entirely one person" or "both of us, at least partly"]. Following this, we asked those who thought of one-sided blame conflicts to think of a mutual blame conflict. The rest of the procedure is described next.

Procedure for all studies. We asked participants to recall an event from their life where they experienced a mutual blame conflict with a work colleague (in Studies 1a and 1b) or anyone from

their life (in Study 1c)—we refer to this person as the target.[1] Participants were presented with six apology scenarios in Studies 1a and 1b and five in Study 1c, and asked to indicate how they would generally feel in each scenario using a sliding scale of -30 = "extremely negative" to 30 = "extremely positive." For the scenarios, we varied whether the participant would be in the position to apologize first or second, and whether one or two apologies occurred in the situation, resulting in four conditions: Both (Self First), Both (Other First), Self Alone, Other Alone. The exact text of these scenarios can be found in Table 2. We also asked about the scenario where neither person apologized (Neither). These five scenarios were common across Studies 1a, 1b and 1c. For testing H1, we focus on the comparison between Self Alone and Both (Self First). To test H2, in Studies 1a and 1b, we also included a scenario in which the participant apologized first, and the other person forgave them (Other Forgives). The first five scenarios were presented in random order across all studies. In Studies 1a and 1b, the Other Forgives scenario, was presented last to prevent participants from being prompted to assume forgiveness was implied in any of the other scenarios.

Table 2. Apology scenarios

Study	**Condition** (within-subjects)	**Scenario**
1a,1b,1c	Self Alone	You apologized first. [Target initials/The other person] did not apologize afterwards.
1a,1b,1c	Both (Self First)	You apologized first, then [target initials/the other person] apologized.
1a,1b,1c	Other Alone	[Target initials/The other person] apologized first. You did not apologize afterwards.
1a,1b,1c	Both (Other First)	[Target initials/The other person] apologized first, then you apologized.
1a,1b,1c	Neither	Neither you nor [target initials/the other person] apologized.
1a,1b	Other Forgives	You apologized first. [Target initials] did not apologize afterwards, but they let you know they forgive you.

Participants. We recruited three different populations during three virtual public talks held over Zoom that consisted of MBA and undergraduate students (Study 1a), international professionals (Study 1b), and members of the general public (Study 1c). Attendees of these three separate talks were asked to spend seven minutes at the beginning of the talk completing a survey through Qualtrics

[1] To prevent participants from lying as a way to continue with the study, in all three studies we allowed participants to continue if they could not think of a scenario from their own life, but could only imagine one. To this end, participants indicated whether they had thought of a real or imaginary conflict where blame was mutual [multiple choice: "I have thought of a real argument/conflict from my life where this is true" or "I am imagining myself in a fictional situation where this is true (I can't think of one from my own life)"].

in exchange for a chance to win a $25 Amazon gift card. Our hypotheses are based on real mutual blame conflicts; hence, we exclude from the analysis participants who could not think of a real conflict where both parties where partly to blame. We further exclude participants who failed the attention check.[2] For Study 1a, 61 unique participants started the study and 48 completed it. We excluded one participant who failed to pass the attention check. Out of the remaining 47 participants, four could not recall a real scenario (the scenarios recalled were imaginary mutual-blame ones) and one could only recall a one-sided blame conflict scenario (though real). Excluding these five participants, leaves a sample of 42 (42.9% male, M_{age} = 34.3 years, SD_{age} = 20.6). For Study 1b, 597 unique participants started the study and 474 completed it. From these, we excluded 139 participants who failed to pass the attention check. Out of the remaining 335 participants, as per our pre-registration, we excluded anyone outside of the first 200 participants who completed the study and passed the attention check. From these, 41 participants could not recall a real scenario and four participants could only recall a one-sided blame conflict scenario. One participant out of these could not do either. Hence, there were 44 participants who could not recall a real and/or mutual-blame conflict. Excluding them, leaves a sample of 156 (50.0% male, M_{age} = 31.8 years, SD_{age} = 6.1). For Study 1c, 108 unique participants started the study and 67 completed it. We excluded six participants who failed to pass the attention check and five who could only recall a one-sided blame conflict scenario (as measured both by the initial binary question unique to this study and by the final continuous blame measure). All the remaining 56 participants (44.6% male, M_{age} = 44.0 years, SD_{age} = 16.9) recalled a real mutual blame conflict.

Results

In Study 1c, when asked to think of any unresolved conflict from their lives, 93% (57/61) recalled a mutual blame conflict. (Recall that those who did not think of a mutual blame conflict initially (N = 4) were then asked to do so.)

[2] Our pre-registration for Studies 1a and 1b mentioned only the passing of the attention check as an exclusionary criterion. Given our hypotheses, we also need to exclude participants who reported a one-sided blame conflict scenario or that they could only imagine one – exclusionary criteria we specify in the pre-registration for Study 1c. Nevertheless, our findings for Studies 1a and 1b are essentially identical when including these two groups of participants for our hypothesis testing. We report this analysis in Appendix B Table B.2.

As can be seen in Figure 2, participants indicated that apologizing first and not getting a return apology would feel, on average, negative (1a: M = -18.76, SD = 12.01; 1b: M = -2.11, SD = 12.67; 1c: M = -19.68, SD = 11.81) and worse than the average positive feelings associated with apologizing first and getting a return apology (1a: M = 6.98, SD = 14.78; 1b: M = 5.08, SD = 11.00; 1c: M = 10.25, SD = 14.30), 1a: $t(41) = 11.5$, $p < 0.001$, $d = 1.78$; 1b: $t(155) = 6.40$, $p < 0.001$, $d = 0.51$; 1c: $t(55) = 14.50$, $p < 0.001$, $d = 1.94$.

For Studies 1a and 1b we also found that people expected that receiving forgiveness in response to their apology would feel worse (1a: M = -12.60, SD = 16.15; 1b: M = 2.24, SD = 11.43) than receiving a return apology, 1a: $t(41) = 6.94$, $p < 0.001$, $d = 1.07$; 1b: $t(155) = 2.79$, $p = 0.006$, $d = 0.22$. See Table B.1 in Appendix B for average feelings in the other apology scenarios and for additional comparisons across scenarios.

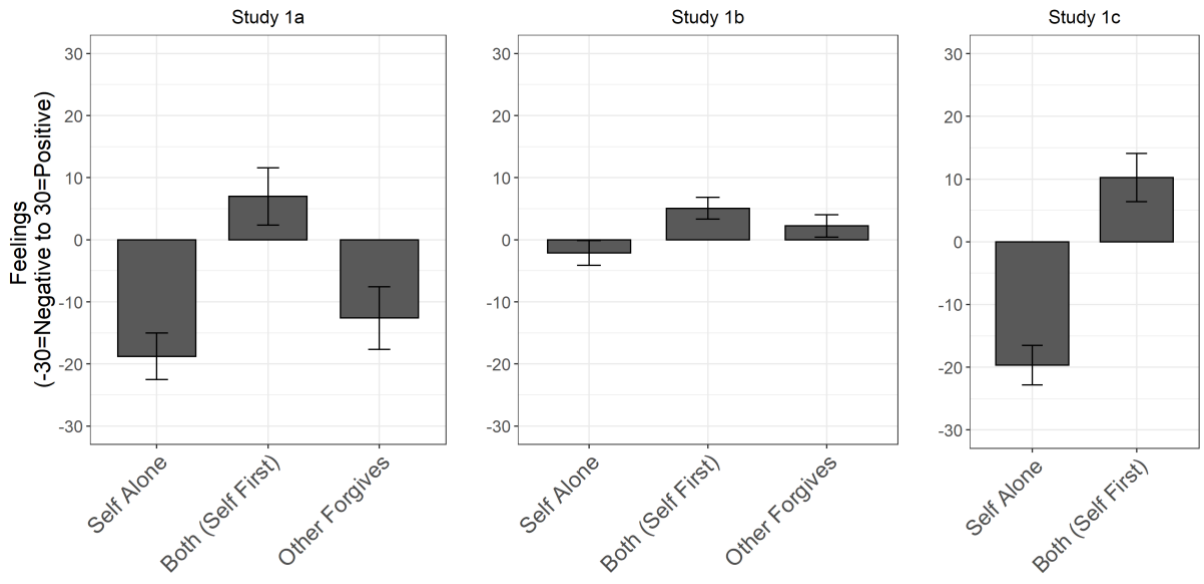


Figure 2. Anticipated feelings for apologizing and not getting a return apology (Self Alone), for apologizing and getting a return apology (Both (Self First)), and for apologizing and receiving forgiveness (Other Forgives) across three different populations (Study 1). Error bars represent 95% confidence intervals.

Discussion

This study finds initial evidence that the majority of unresolved conflicts that people spontaneously recall are mutual blame conflicts. This implies our framework could be of significant value in navigating unresolved conflicts, but given the small sample size of Study 1c, we seek to replicate this result in Study 4 before discussing it further.

This study also confirms our hypothesis that the possible outcomes to apologizing first in a mutual blame conflict are not viewed as having equal value. Across three different samples and a variety of self-reported real-world conflicts, parties in mutual blame conflicts expect to feel better after apologizing and receiving a return apology than after apologizing and not receiving a return apology (H1), even if in the latter case the counterpart offers forgiveness (H2). Furthermore, these outcomes are anticipated to be qualitatively distinct: Receiving a return apology is expected to result in positive feelings, whereas apologizing alone is expected to result in negative feelings.

Despite this, parties in a conflict may still prefer to take the risk of apologizing first if they expect apologizing alone will not feel any worse than the status quo. As described earlier, the worst version of the status quo is when neither person apologizes. As can be seen in the additional analyses in Appendix B, while participants expected to feel directionally worse when apologizing alone than when neither person apologizes (Neither scenario), these differences were not significant, 1a: $t(41) = 1.7$, $p = .099$, $d = 0.26$; 1b: $t(155) = 0.2$, $p = .835$, $d = 0.02$; 1c: $t(56) = 1.3$, $p = .187$, $d = 0.18$. If people do not perceive apologizing alone as worse than the status quo, then the first mover's likelihood to apologize first should not depend on information about the likelihood of getting a return apology—i.e., H3 should not be supported. We test this in the next three studies.

4. Study 2: Conditional initiation in real choices to apologize

Study 2 was designed to test whether the choice to apologize first is sensitive to beliefs about the likelihood of getting a return apology (H3). To this end, we created real mutual blame conflicts (with real monetary incentives) between interacting first mover-second mover dyads, and we manipulated the first mover's perceived likelihood of getting a return apology. To manipulate

perceived likelihood, we conducted the study during the Jewish holiday period of Aseret Yemei Teshuva, a ten-day period between Rosh Hashanah and Yom Kippur in which observant people are encouraged to apologize more than usual. This is an indirect manipulation of likelihood to apologize but a natural one which lends the results to higher generalizability given other such natural cues people may be exposed to. Half of the first movers were paired with a person observing the holiday, and the other half were paired with a nonobservant person. First movers were, as their role implies, given the first opportunity after the conflict occurred to send a free-form message to their partner, and these messages were coded for the presence of apologies. According to H3, we expected that those paired with observant partners would be more likely to apologize than those paired with nonobservant partners. This study's hypotheses and design were pre-registered at https://aspredicted.org/XZW_HKH.

Methods

General procedure. Each pair consisted of one first mover and one second mover, as described in the theory section. These roles corresponded to the order in which participants were able to send a message to their partner after the conflict occurred. Because the study needed to be conducted during a ten-day period for the manipulation to be truthful, and because we expected to need about 300 pairs, we conducted the study asynchronously and matched each second mover to multiple first movers. This was not made explicitly known to first movers. The survey was divided into two parts to allow for asynchronous interactions between first and second movers (see Figure 3 for the timeline of the experiment).

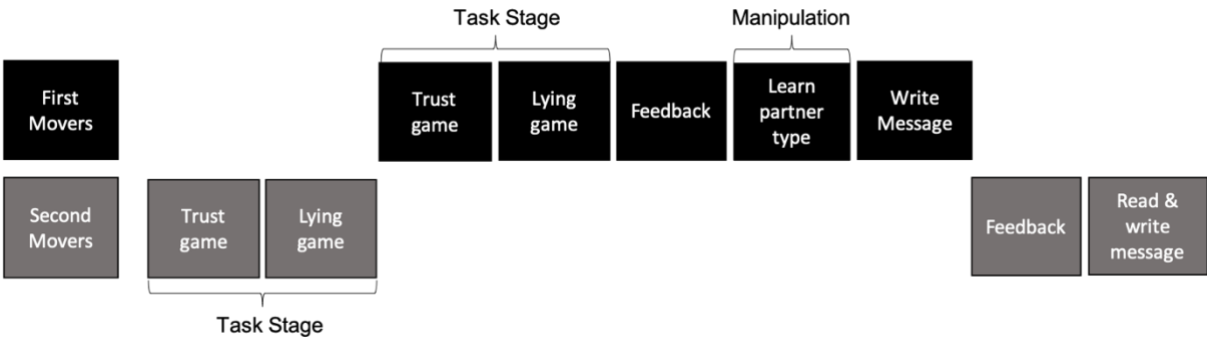


Figure 3. Timeline of experimental procedures for Study 2.

To create a real conflict, each participant in the pair completed a task with two stages, in which each person in the dyad was incentivized to offend their counterpart in one of the two stages. Part one consisted of a trust game and part two consisted of a lying game, both with real monetary stakes. Monetary amounts were converted to chips (500 chips = $1.00). The games were structured to motivate participants to select options that would lead to a mutual blame conflict: First movers were led to make an unfair allocation in the trust game, and second movers were led to choose a deceptive message in the lying game. (For details of these tasks, see the *Trust game procedure* and *Lying game procedure* sections in Appendix C.) After completing both stages of the task, first movers immediately received feedback about each person's choices and earnings.

After learning about the offenses, one person (first mover) was given the option to send a free-form message to their counterpart (second mover), who they knew would then be given the opportunity to respond with a free-form message. As mentioned above, we manipulated perceived likelihood of getting a return apology by conducting the study during the Jewish holiday period of Aseret Yemei Teshuva (which we referred to in the study as the "Ten Days of Apologizing"), a period in which observant people are encouraged to apologize more than usual. We used the name Ten Days of Apologizing to abstract from a religious context in order to minimize the influence of any positive or negative associations first movers might have with religion. Half of the first movers were paired (and told they were paired) with someone observing this holiday period (HIGH condition), and the other half were paired with someone not observing this holiday period (LOW condition).

To create the partner-type manipulation, at the end of the second movers' first survey, they were told the purpose of the study and asked to indicate whether they were observing the Jewish practice of Aseret Yemei Teshuva during September 2020. (We told them we would be referring to it as the "Ten Days of Apologizing.") They were asked to answer the question, "Which of the following best describes you?" by selecting one of two options: "I will NOT be observing the Ten Days of Apologizing, so during September 18-28, 2020 I will NOT be apologizing any more than I would normally"; "I WILL be observing the Ten Days of Apologizing, so during September 18-28, 2020 I WILL be apologizing MORE than I would normally."

Right after getting task feedback, first movers learned that they were taking this study during the Ten Days of Apologizing. They were then shown a screenshot of the question their partner was asked and their partner's answer. (For screenshots of what the manipulation looked like, see Figures C.1 and C.2 in Appendix C.) We included a comprehension check on the page of the manipulation to confirm participants had read and absorbed the manipulation information on the screen. Specifically, first movers indicated which option was selected by their partner.

After learning whether their partner was observant or not, first movers were then given the option to write a message to their partner, knowing their partner would be able to send a message of their own afterwards. Following this, we elicited perceived relative blame by asking, "How much blame do you think each person deserves for the results of the task stage?" The response scale was a slider from 0 = "You deserve 0% of the blame; Your match deserves 100% of the blame" to 100 = "You deserve 100% of the blame; Your match deserves 0% of the blame." After the blame question, first movers answered a series of exploratory questions, as well as a comprehension check and demographics. For details of these measures, see the online supplement.

For our main measure of apology, two research assistants blind to our hypotheses coded the first movers' messages for the presence of apologies. Coders rectified disagreements through discussion.

Participants. Second movers completed both stages of the task before the first movers, and they were only paired with a first mover if they selected the options that would lead to a mutual blame conflict in the expected way. To find second movers for the observant condition, we recruited from a variety of sources including an Israeli Facebook group, a synagogue in the US Midwest, an Israeli university, as well as personal connections. To find second movers for the nonobservant condition, we recruited from a participant pool associated with the behavioral lab at a large midwestern university in the United States. The first survey was completed by 19 observant participants, but only 11 made choices that could create a mutual blame conflict and thus made them eligible to be paired with a first

mover. Of the 24 nonobservant participants who completed the first survey, 11 made decisions that made them eligible to be paired with a first mover.[3]

First movers were recruited from the same behavioral lab participant pool as nonobservant second movers. We pre-registered that we would collect a sample of 300 first movers who passed comprehension checks and met two additional criteria: (1) They ended up in a mutual blame conflict based on their task choices (i.e., a mutual blame conflict would be one where both people in the pair have made a choice in one of the games that negatively impacts their partner), and (2) they perceive both themselves and the other person to deserve at least part of the blame (as measured by the blame question). We obtained 517 first movers who fully completed the study. Of those 429, passed the comprehension check in Part 1, were paired with a second mover, and completed Part 2. Each participant was paired with one of the 22 eligible second movers. Of those pairs, 288 made choices that created a mutual blame conflict between the two individuals. According to the subjective measure of relative blame, within this sample there were eight pairs in which the first mover either claimed all the blame (7) or assigned all the blame to their partner (1), suggesting they did not see it as a mutual blame conflict. The final sample was 280 (24% male, M_{age} = 26.78 years, SD_{age} = 9.15)—140 in both the observant and nonobservant conditions.

First movers received $4.00 for completing the first part of the survey (15-20 minutes) plus up to a bonus of $2.00 based on the outcome of the task. For the second part of the survey (7-10 minutes), first movers were guaranteed $3.00 for completion and up to a $1.00 bonus. For the trust game portion of the experiment, the pair could collectively earn $1.20, and for the lying game portion of the experiment, each person in the pair could earn up to $0.80. To incentivize attention on the first survey, first movers were told they would only receive bonus payments if they passed the comprehension checks. They were paid on average $8.88 ($17.76/hour)

[3] We pre-registered that we would collect 15 second movers per condition, rather than 11. However, because the choices from the second movers is not central to our hypotheses tests, this does not interfere with our results. It only affected the logistics of the experiment, increasing the number of first movers we had to pair with each second mover, which was not known to first movers.

Results

Relative blame was perceived to be similar and approximately 50-50 in both conditions (LOW: M = 48.82, SD = 14.62; HIGH: M = 51.16, SD = 14.33; $t(278) = 1.35$, $p = 0.177$, $d = 0.16$). Consistent with the intention of the manipulation, participants perceived observant partners to be more likely to send a return apology, M = 2.34, SD = 16.84, than nonobservant partners, M = -2.43, SD = 14.69, $t(278) = 2.53$, $p = 0.0121$, $d = 0.30$ (using a -30 = "extremely unlikely" to 30 = "extremely likely" sliding scale). There were no cross-condition differences in the overall number of messages sent by first movers (LOW: 69.29%, 97/140; HIGH: 70.00%, 98/140; X^2 (1, N = 280) = 0.02, $p = 0.8966$, $\Phi = 0.01$; Figure 4).

Based on the coding of two RAs blind to our hypothesis, we found that first movers were directionally more likely to apologize when they were partnered with an observant second mover (25.71%, 36/140) than when they were partnered with a nonobservant second mover (17.14%, 24/140), but this difference was not significant, X^2 (1, N = 280) = 3.05, $p = .0805$, $\Phi = 0.10$ (Figure 3). After relaxing exclusions to include all first movers who experienced a mutual blame conflict as defined by choices (N = 288), the difference is of a similar size and significant at the 5% level: HIGH: 26.57%, 38/143; LOW: 16.55%, 24/145, X^2 (1, N = 288) = 4.28, $p = .0386$, $\Phi = 0.12$. All other relationships between main variables hold (see Appendix C for additional analyses).

Given the small effect size of our pre-registered analysis, we conducted a robustness check to verify the directionality of our main result. To do this, we conducted a separate round of message coding using MTurk workers. Each worker coded 30-40 messages such that each message was coded by 8-10 MTurk workers. Coders were told that messages came from pairs of participants who worked on an online task together. For each message they were asked, "Does the speaker apologize to the listener in the following message?" using a sliding scale from -30 = "definitely not" to 30 = "definitely yes." For each message, we averaged the codes across workers. Messages with positive average values were considered to have an apology whereas those with average score equal or less than zero were considered to not have an apology.

According to this coding measure, first movers were also directionally more likely to apologize in the HIGH condition (22.9%, 32/140) than in the LOW condition (16.4%, 23/140), X^2 (1, N = 280) = 1.83, p = 0.1758, Φ = 0.080 (Figure 4). We found a similar result after relaxing exclusions to include all first movers who experienced a mutual blame conflict as defined by choices: LOW: 15.9%, 23/145; HIGH: 23.1%, 33/143, X^2 (1, N = 288) = 2.39, p = 0.1219, Φ = 0.091.

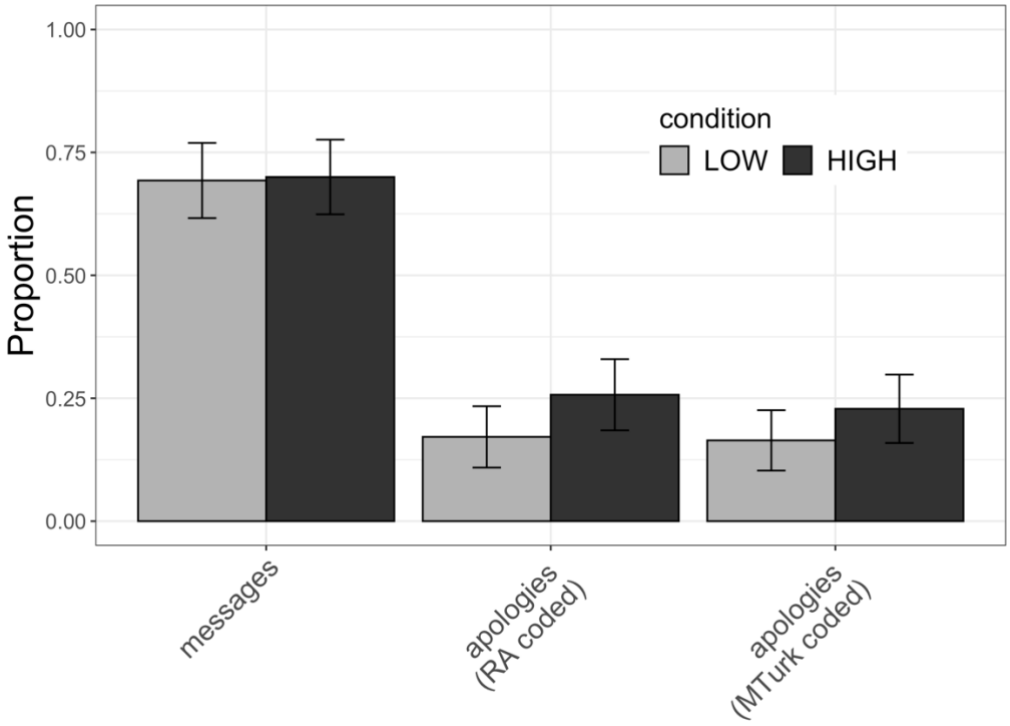


Figure 4. Proportion of first movers in Study 2 who sent messages, proportion who apologized as coded by two research assistants ("RA coded"), and proportion who apologized as coded by MTurkers ("MTurk coded"). Error bars represent 95% confidence intervals. All proportions calculated based on total number of participants within the associated condition.

Discussion

By orchestrating real mutual blame conflicts with real monetary stakes between interacting dyads and allowing partners to write free-form messages to each other, we found suggestive evidence of conditional initiation (H3): First movers were directionally more likely to apologize when they thought their partner was likely, as opposed to unlikely, to apologize in return. However, this result

was not statistically significant. As a result, we checked the robustness of the directional result by conducting a separate round of coding with more coders per message and with a continuous measure of apology. We replicated the directional result.

We chose this design with anonymous interactions and pre-structured conflicts to reduce variance between pairs. This design allowed us to make all pairs experience almost exactly the same conflict, and to avoid influence from individual characteristics of the participants (e.g., gender, attractiveness, speech accent, etc.). However, this may also explain why the effect we detected was smaller than anticipated, such that our sample was not large enough to accurately estimate it (i.e., $\Phi = 0.10$).[4] The fact that the interactions were anonymous and asynchronous likely muted the emotional components of the experience of apologizing, reducing the perceived emotional difference between apologizing alone (i.e., Y in Figure 1) and apologizing together (i.e., X in Figure 1). Furthermore, the benefits of apologizing are primarily in its ability to repair ongoing relationships. At the time of writing their messages, the first movers in this study had no reason to anticipate any future interactions with the second movers. Finally, about 65% of first movers reported believing that neither person needed to apologize (despite over 85% reporting they felt offended by the second mover's behavior)[5]. This may be due to how we structured the task to nudge participants to commit an offense. That is, they may have felt the circumstances justified their behavior.

Given this suggestive evidence for H3, we conducted two additional studies to test H3 and address some of the weaknesses of Study 2. In Studies 3 and 4, we had participants consider existing conflicts outside the lab. In Study 3, participants were faced with the same conflict involving an unresolved dispute at the level of their country. We then asked about their willingness to support their president's willingness to apologize. In Study 4, participants had to recall unresolved conflicts from their own lives. While this latter design means each participant considered a different conflict, it also

[4] A post-hoc power analysis revealed that, with a sample of 280, our power was approximately 39%. Because this study needed to be conducted within the 10-day period of Aseret Yemei Teshuva for our manipulation to be truthful, and because we used a university lab and its participant pool, we were constrained to collect fewer than 300 first movers.

[5] For more details on these measures, see Table S.4 in the online supplement.

incorporated the emotional and reputational concerns that are endemic to most interpersonal conflicts outside the lab.

5. Study 3: Conditional initiation in support for leader apologies in intergroup conflicts

Study 3 was designed to further test whether the choice to apologize first is sensitive to beliefs about the likelihood of getting a return apology (H3). Similar to Study 2, all participants considered the same conflict, however, here participants considered an intergroup (rather than an interpersonal) conflict. In particular, we asked US citizens to imagine a situation where US president Joe Biden decided to apologize to the Japanese prime minister for the bombings of Hiroshima and Nagasaki during World War II (WWII). We tested whether participants would be more supportive of the president's choice when they believed the Japanese prime minister was likely (vs. unlikely) to respond by apologizing for bombing Pearl Harbor. This study's hypotheses and design were pre-registered at https://aspredicted.org/4J3_L9N.

Methods

Procedure. Participants were asked to imagine that the US president was going to meet with the Japanese prime minister the following week when they would both give condolences to the victims of the bombings of Pearl Harbor and Hiroshima and Nagasaki. However, it was not clear whether either side would make an explicit apology on behalf of their nation. In the CONTROL condition, there was no additional text. In the LOW (HIGH) condition, participants also read, "Earlier last week, Prime Minister Yoshihide Suga declared in an interview that he would not apologize (would apologize) for Japan's bombing of Pearl Harbor, even if (if) the US President, Joe Biden, first apologized for the bombing of Hiroshima and Nagasaki." We elicited participants' support for US apologizing first by asking the following question: "Imagine that President Biden decided he was going to initiate an apology (i.e. apologize first) to Japan's Prime Minister Yoshihide Suga, for the WWII bombings of Hiroshima and Nagasaki during next week's meeting. Would you support President Biden's choice?". Participants provided their answer using a sliding scale from -30 = "definitely no" to 30 = "definitely yes." Participants indicated perceived relative blame for the US and

Japan separately by responding to the following question: "How much blame do you think each country deserves regarding the combined deaths in Pearl Harbor, Hiroshima, and Nagasaki during the bombings of World War II?" Participants utilized two sliding scales from 0 = "no blame" to 100 = "all of the blame." Perceived relative blame was calculated by taking the amount of blame assigned to the US divided by the sum of the blame assigned to the US and Japan.

Participants. We recruited participants through Prolific and screened for US citizenship. Participants completed a 5-8 minute survey for $0.80 base pay plus $0.20 bonus for correctly completing the comprehension checks. Of the 1041 unique participants that started the survey, 921 passed the attention check, completed the survey, and answered the comprehension check question correctly. As we pre-registered, our main analysis was done on the subsample of the first 300 participants in each condition who reported a mutual blame conflict (each conflictual party had some positive non-zero blame) in the scenario they were presented with, leaving 899 US citizens (41% male, M_{age} = 34.14 years, SD_{age} = 12.20).

Results

Participants perceived the scenario used in the study as involving a mutual blame conflict since the distribution of relative blame was perceived to be approximately 50-50 in both conditions (LOW: M = 45.68, SD = 16.77; HIGH: M = 46.04, SD = 14.54; $t(586.24) = -0.28$, $p = 0.780$, d = 0.02). Moreover, our manipulation was successful in leading participants to perceive a lower likelihood that the US would receive a return apology from Japan's prime minister in the LOW condition, M = -20.41, SD = 11.71, than in the HIGH condition, M = 20.55, SD = 12.45, $t(598) = -41.51$, $p < 0.001$, d = 3.39. In the CONTROL condition, participants' mean perception was in between the LOW and HIGH conditions, M = 8.81, SD = 15.11. This is an indication that our manipulation moved participants' perceptions beyond their prior beliefs (LOW vs. CONTROL: $t(561.20) = -26.46$, $p < 0.001$, d = 2.16; HIGH vs. CONTROL: $t(575.27) = 10.38$, $p < 0.001$, d = 0.85).

In line with H3, participants' support for the US president's apology was sensitive to beliefs that Japan would return the gesture (Figure 5): They were more likely to support the president's initiating an apology for the US's past behavior towards Japan in WWII when they thought it was more likely that Japan's prime minister would apologize in return for Japan's past behavior towards

the US in WWII, M = 14.41, SD = 18.64, than when they thought this was less likely to happen, M = 8.62, SD = 19.56, $t(598) = -3.71$, $p < 0.001$, $d = 0.30$.

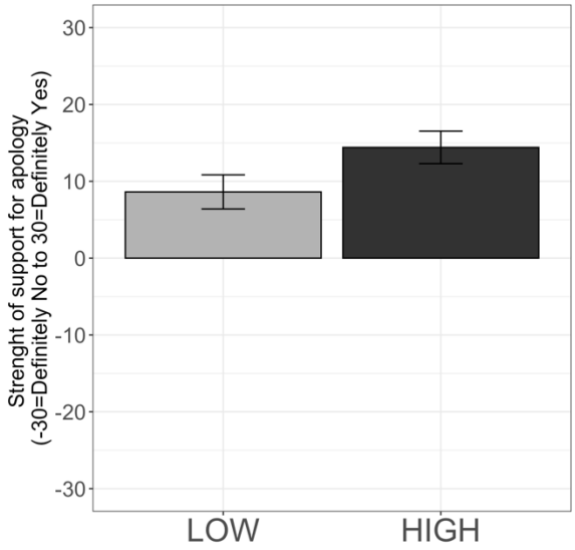


Figure 5. Average strength of support for US president's choice to initiate an apology to the Japanese prime minister for bombings the US committed in WWII (Study 3). Error bars represent 95% confidence intervals.

Discussion

Study 3 confirmed that people are sensitive to the likelihood of getting a return apology even for group level apologies. In particular, information about a return apology impacted their support for their group leader's choice to apologize to another group on their behalf. Apologies on behalf of groups play an important role in diplomatic communication, transitional justice and international as well as intranational reconciliation (Barkan & Karn, 2006; Kitagawa & Chu, 2021; Marrus, 2007). The person making an apology on behalf of a group is usually a group leader of some sort, be it the CEO of an organization or the president of a country. Whether the followers support the leader's choice to apologize or not could be consequential for that leader's tenure. Our findings suggest that conditional initiation is a dynamic that is important for leaders to consider in their choice to apologize on behalf of their followers.

6. Study 4: Conditional initiation in recalled conflicts

Our goals in Study 4 were three-fold. First, we wanted to supplement our findings from Study 1c by collecting another observation assessing the proportion of unresolved conflicts that are mutual blame, as opposed to one-sided blame. Second, we aimed to examine whether sensitivity to getting a return apology (H3) is present in real interpersonal conflicts outside the lab. Third, we aimed to test whether interactants are more likely to apologize second, after having received an apology from their counterpart, than they are to give the initial apology (H4). To this end, we had participants recall unresolved conflicts from their lives. This study's hypotheses and design were pre-registered at https://aspredicted.org/VKZ_UTW.

Methods

Procedure. To assess the share of unresolved conflicts that involve mutual blame, we conducted a similar procedure to the initial procedure for Study 1c: We first asked participants to think of any unresolved conflict from their life. If they could not think of such a conflict, they were asked to return the submission. Participants indicated who was to blame for the conflict [multiple choice: "entirely one person" or "both of us, at least partly"], and then indicated how much blame each person deserved for the conflict using a similar continuous measure to that in Study 2. Following this, we asked those who thought of one-sided blame conflicts to think of a mutual blame conflict. They then answered the two blame questions for this conflict. After this, participants indicated whether either person offered an apology to the other, regardless of whether the apology was effective, complete, or sincere [multiple choice: "only one person apologized (either me or [initials of other])," "both of us apologized," "neither of us apologized").

Participants who indicated that neither person had yet apologized were our target group for testing H3 and H4, so only those participants were asked the questions described next. To assess whether their beliefs about the likelihood of getting a return apology would affect their willingness to initiate an apology, participants were presented with two scenarios (within-subjects). The scenario meant to give low expectations of a return apology (LOW condition) was worded as follows: "Now, imagine that you learned through a third party that [initials of other] was very unwilling to apologize

to you for their part in the conflict, suggesting they probably will not apologize at any point. In this scenario, how likely would you be to apologize first?" The scenario meant to give high expectations of a return apology (HIGH condition) was worded the same except "unwilling" and "will not apologize at any point" were replaced with "willing" and "will apologize at some point." Both of these were answered using a sliding scale of -30 = "extremely unlikely" to 30 = "extremely likely." The order of these scenarios was counterbalanced.

The target group was also presented with two questions assessing their likelihood to apologize first versus second (within-subjects): "If you have the chance, how likely are you to apologize to [initials of other] first?" (FIRST condition) and "Imagine that [initials of other] offers an apology to you. How likely would you be to apologize to them afterwards?" (SECOND condition). Both questions were answered using a sliding scale of -30 = "extremely unlikely" to 30 = "extremely likely. The order of these questions was counterbalanced.

We also asked a series of exploratory questions to the target group as well as to the other participants. See the online supplement for screenshots of the survey and a list of these additional measures.

Participants. We recruited participants through Prolific. Participants completed a 7-12 minute survey for $0.70 base pay plus $0.70 bonus for correctly completing the comprehension checks. Of the 381 unique participants that started the survey, 283 passed the attention check, completed the survey, and answered the comprehension check question correctly (45% male, M_{age} = 31.7 years, SD_{age} = 10.7). As pre-registered, our main analysis was done on the subsample of the first 150 participants who reported a mutual blame conflict in which neither they nor the other person had yet apologized (38% male, M_{age} = 32.6, SD_{age} = 11.3).

Results

When asked to think of any unresolved conflict from their lives, 63% (178/283) recalled a mutual blame conflict. Of these, 58% (104/178) were situations where neither person had yet apologized. Recall that those who did not think of a mutual blame conflict initially (N = 105) were then asked to do so. As pre-registered, we tested H3 and H4 using the first 150 of these participants who recalled a mutual blame conflict where neither person apologized.

To test for conditional initiation (H3), as pre-registered, we used a mixed-effects linear regression, with likelihood to apologize first as the dependent variable regressed on a binary variable for the manipulated scenario (0 = LOW, 1 = HIGH). We controlled for the order in which the manipulated scenarios were presented (including a main effect and interaction with the binary scenario variable), as well as for perceived relative blame and participant-level heterogeneity (using random effects; *Pseudo-*$R^2_{\text{fixed effects}}$ = 0.39). We found support for H3 (Figure 6): In real life conflicts, participants reported they would be less likely to apologize first when they thought the other person would be unlikely (LOW: M = -23.87, SD = 12.37) compared to likely to return the apology (HIGH: M = -4.14, SD = 20.80), $t(148.00) = 9.53$, $p < 0.001$, $\beta = 1.08$. A full regression table can be found in Table D.1 in Appendix D.

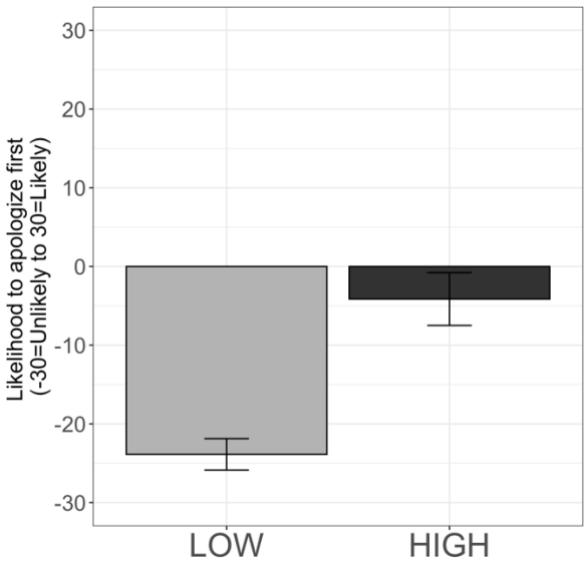


Figure 6. Average likelihood to apologize first in recalled conflicts (Study 4). Error bars represent 95% confidence intervals.

To examine whether participants were more likely to apologize second than first in the real life conflicts they reported (H4), as pre-registered, we used a mixed-effects linear regression, with likelihood to apologize as the dependent variable regressed on a binary variable for order of apology (0 = FIRST, 1 = SECOND). While controlling for participant-level heterogeneity by using random effects (*Pseudo-*$R^2_{\text{fixed effects}}$ = 0.30), we found that participants reported being less likely to apologize

first (M = -15.87, SD = 17.65) than second (M = 9.29, SD = 21.08), in conflicts they faced in their past where neither person had yet apologized, $t(149.00) = 15.19$, $p < 0.001$, $\beta = 1.09$ (Figure 7). This result did not change after controlling for perceived relative blame and the order in which participants answered the questions about the likelihood to apologize first and second (see Table D.2 in Appendix D for a full regression table).

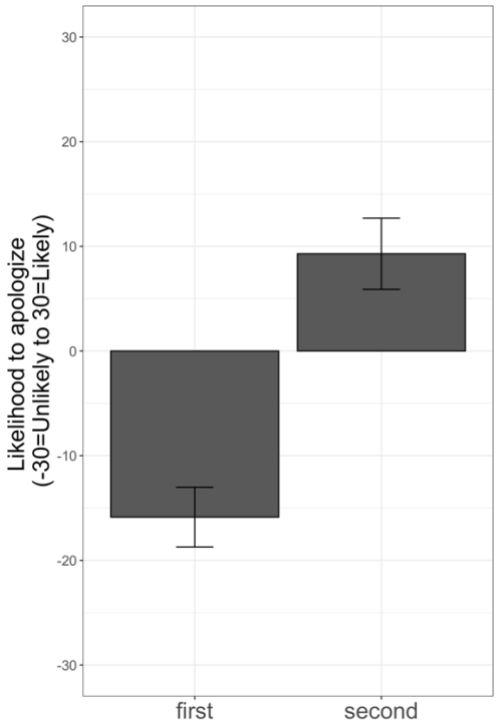


Figure 7. Self-reported likelihood to apologize first versus second in recalled mutual blame conflicts (Study 4). Error bars represent 95% confidence intervals.

Discussion

Study 4 found that mutual blame conflicts constitute a large portion of the unresolved conflicts that people recall, and that within these conflicts, the majority of interactants have not yet apologized to one another. This highlights the need to better understand the reconciliatory processes within such situations. Our framework constitutes a step in that direction.

Study 4 also provided evidence that, in real interpersonal mutual blame conflicts, people's willingness to apologize first is sensitive to whether they think the other person is likely to apologize

or not (H3), and people are more likely to apologize second than first (H4). Both of these hold even after controlling for perceived relative blameworthiness.

While people may self-report on a continuous scale that they would be more willing to apologize second than first (as in Study 4), this may not be true when people are making their decisions more implicitly, as in real life when they are choosing among a multitude of discrete messages in their head. Furthermore, because we asked about the willingness to apologize first and second within subjects in Study 4, participants could have been directly comparing their answers and responding accordingly. With the clear counterfactual of apologizing first, people may prefer to apologize second. But what happens when people are merely confronted with either the choice to apologize first or the choice to apologize second? Do they still show a greater willingness to apologize second? We address these issues in Study 5 by manipulating message order between subjects (rather than within subjects) and by providing participants with a set of six discrete message options to select from, only one of which contains an apology.

7. Study 5: The impact of order on the likelihood of apologizing

The aim of Study 5 was to examine whether participants were more likely to apologize second than first (H4) in the context of a between subjects manipulation of apology order and also using discrete message choices as the measure of willingness to apologize rather than a continuous scale. This allows for a more indirect, less explicit measure of the likelihood to apologize. We also use a scenario design to control features of the conflict between subjects. The idea for the scenario came from one of the real recalled conflicts in Study 1a: Two work colleagues were responsible for a mistake made on a project because, while one person was responsible for creating the work and producing the mistake, the other was responsible for checking for mistakes. We had participants imagine themselves in an embellished version of this scenario and make choices for how to converse with their colleague after a mistake. We expected participants would be more likely to select an

apology message, if the colleague apologized first.[6] This study's hypotheses and design were pre-registered at https://aspredicted.org/P6Q_3CR.

Methods

Procedure. Participants were asked to think of a very close friend from their real life and type out their initials to be used in the scenario as the other person. We did this to increase participants' ability to imagine and care about the other person in the scenario. We then asked participants to imagine themselves and this person in a workplace scenario. In the scenario, the participant learned they and their friend worked at a financial consulting firm and both were assigned to work on a project together. The participant was assigned to generate a report, and the friend was assigned to check the report for mistakes before sending it to the client. After sending the report to the client, the client emailed that they had found a major error and, as a result, decided to end their relationship with the consulting firm. The participant also learned their supervisor was furious and was considering firing them. At this point, we had participants answer three comprehension check questions to ensure they understood the details.

Participants then imagined they ran into their colleague in the breakroom before they were both scheduled to go to their supervisor's office to discuss the mistake. Half the participants were told the colleague simply said, "Hey, how are you?" (FIRST condition) while the other half were told the colleague said the same thing but followed it with an apology, "Hey, how are you? I'm really sorry for not catching that error! I feel terrible" (SECOND condition). Participants were asked what they would say in response and were given six message options to choose from: an apology message, a blaming message, a forgiving message, and three neutral, non-attributional messages (presented in random order). They could select as many messages as they wanted. See Figure E.1 for a screenshot of this question, and see Appendix E for the exact wording of each message for each condition. Perceived relative blame was assessed using a similar measure and calculation as that in Study 3. We also collected demographics.

[6] We also planned to test whether agents' choice to apologize first was sensitive to the magnitude of outcomes, i.e., how bad they expect to feel if they end up apologizing alone. However, the data did not meet the assumptions of the test we planned to use. For the sake of brevity, we moved discussion and analysis of this hypothesis to Appendix E.

Participants. We recruited participants through Prolific. Participants completed an 8-15 minute survey for $1.50 base pay plus $0.50 bonus for correctly completing the comprehension checks. Of the 1030 unique participants that started the survey, 807 (407 in the first-mover condition; 400 in the second-mover condition) passed the attention check, completed the survey, and answered the comprehension check questions correctly. We pre-registered that we would only collect 400 observations per condition, so we dropped the last seven observations from the first-mover condition. The total sample analyzed was 800 (41% male, M_{age} = 33.8 years, SD_{age} = 12.2).

Results

Participants perceived the scenario used in the study as involving a mutual blame conflict since the distribution of perceived relative blame to the self was approximately 50-50 in both conditions (FIRST: M = 53.62, SD = 13.17; SECOND: M = 55.81, SD = 13.36; $t(798) = -2.33$, p = 0.020, d = 0.16). Participants who received an apology first (SECOND condition), perceived themselves to be slightly more blame than those who had not received an apology (FIRST condition). We control for perceived relative blame in our main analysis.

Using a linear regression and controlling for perceived relative blame to the self (R^2 = 0.13, $F(2,797) = 61.29$, $p < 0.001$), we find that participants are more likely to include the apology message when they are responding to an apology from their colleague (68.25%) than when they have not yet received one (39.75%), $t(797) = 8.07$, $p < 0.001$, $\beta = 0.53$ (Figure 8). See Table E.5 in Appendix E for a table of regression coefficients.

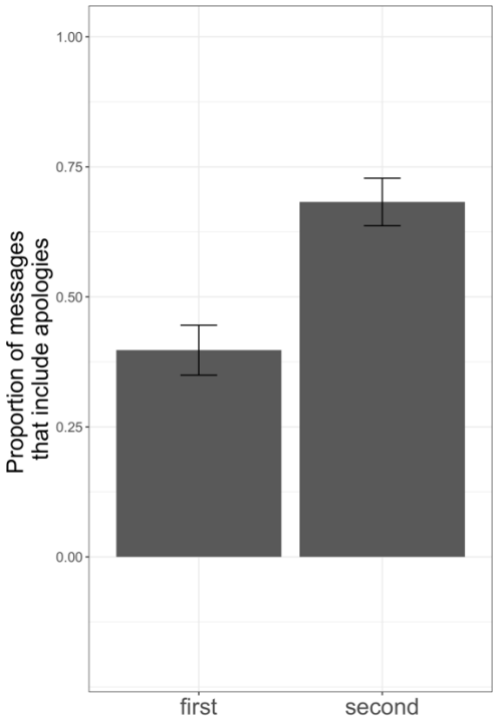


Figure 8. Proportion choosing to apologize first versus second in a workplace scenario (Study 5). Error bars represent 95% confidence intervals.

Discussion

These results demonstrate that people are more likely to apologize second than first (H4), even when they are not directly comparing their willingness to apologize first versus second and even when they are selecting among a set of discrete messages. Importantly, these results hold even after controlling for perceived relative blameworthiness and even though at least 81% of participants in both conditions perceived themselves to be either equally or more to blame than the other person. That is, the preference to apologize when moving second in this study does not stem from a perception that the other person is more to blame, and therefore, should apologize first. While we argue that one reason for this difference is that apologizing first presents a riskier choice, an additional reason for this could be that the desire to reciprocate is present for the second mover but not the first mover. Thus, not only is apologizing second less risky, but there is an additional impetus (i.e., reciprocity) to return the apology. We discuss this possibility and its consequences more in the General Discussion.

8. General Discussion

This paper contributes to the research on apologies and conflict resolution in two important ways. The first contribution is conceptual. We differentiate between unilateral and mutual blame conflicts and focus on the latter which has been unexplored in the apology literature. As we demonstrated in Studies 1c and 4, mutual blame conflicts constitute the largest portion of unresolved conflicts that participants spontaneously recall, highlighting the potential value of research examining reconciliation processes for mutual blame conflicts. Furthermore, we use a game theoretic framework to characterize the strategic decision to apologize in mutual blame conflicts. We model this decision as a sequential game where a first mover decides whether to initiate an apology and a second mover whether to reply in kind. This allows us to identify the elements of this decision-making environment that may influence people's apology behavior: anticipated feelings in and perceived likelihoods of possible outcomes from apologizing. While prior work has identified barriers to apologizing when blame is one-sided, the framework we put forth in this article highlights an additional barrier that is only present when blame is mutual: the risk of not receiving a return apology. This can give rise to conditional initiation of apologies, wherein decision makers' likelihood to apologize first is sensitive to their beliefs about receiving a return apology, and it can leave decision makers waiting to receive an apology before giving one themselves.

Our second main contribution is empirical. We document the characteristics of the risky strategic decision to apologize across multiple contexts and experimental designs. Specifically, we find that the decision to apologize is indeed perceived as a risky choice as people expect to feel worse when apologizing alone than when apologizing and getting a return apology. Moreover, increasing people's perceived likelihood of receiving a return apology increases their likelihood to initiate an apology. Further corroborating the existence of this risk, we find that decision makers are more likely to apologize second (when the risk of apologizing alone is eliminated) than first. Given that our results show that mutual blame conflicts are common, the unique risk associated with apologizing in such instances should be considered as an important source of the reluctance to apologize, and thus, of delayed, or even failed, conflict resolution.

Facing this risk of apologizing first, what should decision makers do? Making the right choice requires them to be accurate in assessing several features of the situation including their counterpart's willingness to reciprocate with an apology. Given that people are particularly bad at perspective-taking (see Epley, 2008), it is likely that decision makers' beliefs are often miscalibrated about their counterpart's willingness to reciprocate an apology, leading them to suboptimal outcomes. To examine this, we conducted a pre-registered scenario study in which participants imagined themselves as one of two individuals in an interpersonal conflict (see Supplemental Study 1 in the online supplement for details). We randomly assigned people to one of the two roles and asked them to report not only their own likelihood to return an apology, but also to guess the other person's likelihood to return an apology. We found that people underestimated the other person's likelihood to return an apology, $t(373) = 5.67$, $p < .001$, $d = 0.29$. If this excessive pessimism about others is pervasive, we expect that people will sometimes end up in *suboptimal* apology stalemates, wherein neither person is willing to initiate an apology, though each would prefer to apologize conditional on getting a return apology. That is, they will be unable to coordinate on the mutually optimal outcome. Such relationships might deteriorate unnecessarily but for one of the two parties involved in the conflict being convinced to make the first overture. A simple solution to this could be for one person to engage in information search about their counterpart's feelings of remorse, for example, by asking indirect questions to their counterpart or by having a mutual friend, therapist, or human resources specialist serve the role of "apology broker," confidentially assessing each person's likelihood to apologize and facilitating apologies when both seem amenable to doing so.

Our framework also makes clear that, to make the best choice, decision makers need to be able to accurately forecast how they themselves will feel in each of the possible resulting situations—i.e., to accurately guess the potential payoffs associated with the decisions to apologize or not. Prior work has found that people make systematic errors in forecasting how they will feel in future situations (see Wilson & Gilbert, 2003), even with respect to apologies (Leunissen et al., 2014). For instance, if people overestimate how bad they will feel when apologizing alone, they will misestimate the risk of apologizing first and be overly reluctant to doing so, diminishing the chance of

reconciliation. One solution to this could be to have individuals engage in self-affirmation which may lower the (mis)perceived ego-threat of apologizing alone (e.g., Schumann, 2014).

The central finding in this paper arises from a consideration unique to mutual blame conflicts: order of apology. While we have focused on how order affects the likelihood of apologizing, this framework predicts that order should have other important consequences that could be the focus of future work. One example is its possible impact on the perceived sincerity of the apology. Perceived sincerity of an apology is a key factor leading to forgiveness (Schumann, 2012; Skarlicki, Folger, & Gee, 2004; Tomlinson et al., 2004; Zechmeister, Garcia, Romero, & Vas, 2004), especially in the eyes of observers (Risen & Gilovich, 2007). According to signaling theory, costlier signals are more sincere, or honest (Zahavi, 1975). Because apologizing first is a riskier (and thus, costlier) action than apologizing second, this framework predicts that the first mover's apology will be perceived as more sincere than the second mover's. Magnifying this, the first apology invokes a norm of reciprocity (Gouldner, 1960), possibly leading the second apology to appear obligatory (Flynn & Yu, 2021). As a first test of this, we ran a study asking participants to read a scenario in which two people had a conflict and apologized to one another (with one of the two going first). Participants judged the sincerity of each apology, and results confirm our prediction: The second apology was rated as less sincere than the first one, $t(76) = 7.64$, $p < .001$, $\beta = 0.73$ (see Supplemental Study 2 for in the online supplement for details). This suggests that another potential solution for the reluctance to initiate apologies could highlight the added sincerity benefits of apologizing first. This could be especially useful in settings with repeated interactions where trust-building is paramount such as in workplace relationships between colleagues.

Signaling theory can also help explain how apology order might impact the perceived blameworthiness of the apologizer, one of the central costs to apologizing. Recall that the risk of apologizing first is not receiving a return apology, a situation that could portray the lone apologizer as fully (rather than partly) to blame. The magnitude of this risk is lower for the counterpart who is actually more to blame—there is a smaller difference between the amount of blame they deserve and full blame. Accordingly, the more blameworthy party should be more willing to apologize first than the less blameworthy party—a prediction we found evidence for in Supplemental Study 1, $X^2(1, N =$

290) = 138.3, $p < .001$, $\Phi = 0.70$. As a result, observers may perceive the person who apologizes first to be more blameworthy than the person who apologizes second (particularly, if the apology content is ambiguous about blame distribution), and indeed, we find support for this in Supplemental Study 2, $t(101.2) = 6.20$, $p < .001$, $d = 1.01$. This reputation threat may create a further barrier to apologizing first, particularly for people who consider themselves less blameworthy than their counterpart.

This work represents an initial investigation into the dynamics of apologizing in mutual blame conflicts. Using a variety of empirical evidence, we have illustrated the existence of strategic concerns which are exclusive to mutual blame conflicts where both conflictual parties can apologize in a sequential order. As such, we have identified additional barriers to apologizing, which can give rise to new solutions for conflict resolution for mutual blame conflicts, which appear to constitute the majority of unresolved conflicts. We have also discussed how the game theoretic setup outlined above can be utilized to think about the consequences of mutual blame beyond the findings we present here, highlighting a new avenue for research in conflict management.

References

Andersson, L. M., & Pearson, C. M. (1999). Tit for Tat? The Spiraling Effect of Incivility in the Workplace. *Academy of Management Review*, *24*(3), 452–471. https://doi.org/10.5465/amr.1999.2202131

Aquino, K., Tripp, T. M., & Bies, R. J. (2006). Getting even or moving on? Power, procedural justice, and types of offense as predictors of revenge, forgiveness, reconciliation, and avoidance in organizations. *Journal of Applied Psychology*, *91*(3), 653–668. https://doi.org/10.1037/0021-9010.91.3.653

Barkan, E., & Karn, A. (2006). *Taking Wrongs Seriously: Apologies And Reconciliation*. *Stanford University Press*.

Baumeister, R. F., Stillwell, A., & Wotman, S. R. (1990). Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger. *Journal of Personality and Social Psychology*, *59*(5), 994–1005. https://doi.org/10.1037/0022-3514.59.5.994

Bies, R. J., & Tripp, T. M. (1996). Beyond Distrust: "Getting Even" and the Need for Revenge. In R.

M. Kramer & T. Tyler (Eds.), *Trust in Organizations: Frontiers of Theory and Research* (pp. 246–260). Thousand Oaks, CA: SAGE Publications Inc. https://doi.org/10.4135/9781452243610.n12

Chaudhry, S. J., & Loewenstein, G. (2019). Thanking, Apologizing, Bragging, and Blaming: Responsibility Exchange Theory and the Currency of Communication. *Psychological Review*. https://doi.org/10.1037/rev0000139

Crawford, V. P., & Sobel, J. (1982). Strategic Information Transmission. *Econometrica*, *50*(6), 1431–1451. https://doi.org/10.2307/1913390

Darby, B. W., & Schlenker, B. R. (1982). Children's reactions to apologies. *Journal of Personality and Social Psychology*, *43*(4), 742–753. https://doi.org/10.1037/0022-3514.43.4.742

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. https://doi.org/10.1111/j.2044-8309.1989.tb00879.x

De Freitas, J., Thomas, K., DeScioli, P., & Pinker, S. (2019). Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences*, *116*(28), 13751–13758. https://doi.org/10.1073/pnas.1905518116

Echterhoff, G., Higgins, E. T., & Levine, J. M. (2009). Shared Reality: Experiencing Commonality with others' Inner States about the World. *Perspectives on Psychological Science*, *4*(5), 496–521. https://doi.org/10.1111/j.1745-6924.2009.01161.x

Epley, N. (2008). Solving the (Real) Other Minds Problem. *Social and Personality Psychology Compass*, *2*(3), 1455–1474. https://doi.org/10.1111/j.1751-9004.2008.00115.x

Feeney, J. A., & Hill, A. (2006). Victim-perpetrator differences in reports of hurtful events. *Journal of Social and Personal Relationships*, *23*(4), 587–608. https://doi.org/10.1177/0265407506065985

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

Flynn, F. J., & Yu, A. (2021). Better to give than reciprocate? Status and reciprocity in prosocial exchange. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspi0000349

Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Review*, *95*(1), 384–394. https://doi.org/10.1257/0002828053828662

Gouldner, A. W. (1960). The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review*, *25*(2), 161. https://doi.org/10.2307/2092623

Greco, L. M., Whitson, J. A., O'Boyle, E. H., Wang, C. S., & Kim, J. (2019). An eye for an eye? A meta-analysis of negative reciprocity in organizations. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000396

Higgins, E. T., & Pittman, T. S. (2008). Motives of the Human Animal: Comprehending, Managing, and Sharing Inner States. *Annual Review of Psychology*, *59*(1), 361–385. https://doi.org/10.1146/annurev.psych.59.103006.093726

Howell, A. J., Dopko, R. L., Turowski, J. B., & Buro, K. (2011). The disposition to apologize. *Personality and Individual Differences*, *51*(4), 509–514. https://doi.org/10.1016/j.paid.2011.05.009

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1111/j.1558-5646.2011.01232.x

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., 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

Kirchhoff, J., Wagner, U., & Strack, M. (2012). Apologies: Words of magic? The role of verbal components, anger reduction, and offence severity. *Peace and Conflict: Journal of Peace Psychology*, *18*(2), 109–130. https://doi.org/10.1037/a0028092

Kitagawa, R., & Chu, J. A. (2021). The Impact of Political Apologies on Public Opinion. *World Politics*, *73*(3), 441–481. https://doi.org/10.1017/S0043887121000083

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807. https://doi.org/10.1037/a0019688

Leunissen, J. M., De Cremer, D., & Folmer, C. P. R. (2012). An instrumental perspective on apologizing in bargaining: The importance of forgiveness to apologize. *Journal of Economic Psychology*, *33*(1), 215–222. https://doi.org/10.1016/j.joep.2011.10.004

Leunissen, J. M., De Cremer, D., van Dijke, M., & Folmer, C. P. R. (2014). Forecasting Errors in the Averseness of Apologizing. *Social Justice Research*, *27*(3), 322–339. https://doi.org/10.1007/s11211-014-0216-4

Lewicki, R. J., Polin, B., & Lount, R. B. (2016). An Exploration of the Structure of Effective Apologies. *Negotiation and Conflict Management Research*, *9*(2), 177–196. https://doi.org/10.1111/ncmr.12073

Lewis, J. T., Parra, G. R., & Cohen, R. (2015). Apologies in Close Relationships: A Review of Theory and Research. *Journal of Family Theory & Review*, *7*(1), 47–61. https://doi.org/10.1111/jftr.12060

Marrus, M. R. (2007). *Official Apologies and the Quest for Historical Justice*. *Journal of Human Rights* (Vol. 6). https://doi.org/10.1080/14754830601098402

Mikula, G., Athenstaedt, U., Heschgl, S., & Heimgartner, A. (1998). Does it only depend on the point of view? Perspective-related differences in justice evaluations of negative incidents in personal relationships. *European Journal of Social Psychology*, *28*(6), 931–962. https://doi.org/10.1002/(SICI)1099-0992(1998110)28:6<931::AID-EJSP904>3.0.CO;2-5

Okimoto, T. G., Wenzel, M., & Hedrick, K. (2013). Refusing to apologize can have psychological benefits (and we issue no mea culpa for this research finding). *European Journal of Social Psychology*, *43*(1), 22–31. https://doi.org/10.1002/ejsp.1901

Pearson, C. M., Andersson, L. M., & Wegner, J. W. (2001). When Workers Flout Convention: A Study of Workplace Incivility. *Human Relations*, *54*(11), 1387–1419. https://doi.org/10.1177/0018726704048355

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(3), 833–838.

https://doi.org/10.1073/pnas.0707192105

Risen, J. L., & Gilovich, T. (2007). Target and observer differences in the acceptance of questionable apologies. *Journal of Personality and Social Psychology*, *92*(3), 418–433. https://doi.org/10.1037/0022-3514.92.3.418

Rossignac-Milon, M., Bolger, N., Zee, K. S., Boothby, E. J., & Higgins, E. T. (2021). Merged minds: Generalized shared reality in dyadic relationships. *Journal of Personality and Social Psychology*, *120*(4), 882–911. https://doi.org/10.1037/pspi0000266

Scher, S. J., & Darley, J. M. (1997). How Effective Are the Things People Say to Apologize? Effects of the Realization of the Apology Speech Act. *Journal of Psycholinguistic Research*, *26*(1), 127–140. https://doi.org/10.1023/a:1025068306386

Schumann, K. (2012). Does love mean never having to say you're sorry? Associations between relationship satisfaction, perceived apology sincerity, and forgiveness. *Journal of Social and Personal Relationships*, *29*(7), 997–1010. https://doi.org/10.1177/0265407512448277

Schumann, K. (2014). An affirmed self and a better apology: The effect of self-affirmation on transgressors' responses to victims. *Journal of Experimental Social Psychology*, *54*, 89–96. https://doi.org/10.1016/j.jesp.2014.04.013

Schumann, K. (2018). The Psychology of Offering an Apology: Understanding the Barriers to Apologizing and How to Overcome Them. *Current Directions in Psychological Science*, *27*(2), 74–78. https://doi.org/10.1177/0963721417741709

Shnabel, N., & Nadler, A. (2008). A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *Journal of Personality and Social Psychology*, *94*(1), 116–132. https://doi.org/10.1037/0022-3514.94.1.116

Skarlicki, D. P., Folger, R., & Gee, J. (2004). When Social Accounts Backfire: The Exacerbating Effects of a Polite Message or an Apology on Reactions to an Unfair Outcome. *Journal of Applied Social Psychology*, *34*(2), 322–341. https://doi.org/10.1111/j.1559-1816.2004.tb02550.x

Tavuchis, N. (1991). *Mea Culpa: A Sociology of Apology and Reconciliation*. Stanford, CA: Stanford University Press.

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

Wilson, T. D., & Gilbert, D. T. (2003). Affective Forecasting. In *Advances in experimental social psychology* (Vol. 35, pp. 345–411). https://doi.org/10.1016/S0065-2601(03)01006-2

Zahavi, A. (1975). Mate selection-A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214. https://doi.org/10.1016/0022-5193(75)90111-3

Zechmeister, J. S., Garcia, S., Romero, C., & Vas, S. N. (2004). Don't Apologize Unless You Mean It : a Laboratory Investigation of Forgiveness and Retaliation. *Journal of Social and Clinical Psychology*, *23*(4), 532–564.

Appendix A. General

OSF page

All data, code, and survey files for all studies can be found on our Open Science Framework page:

https://osf.io/jbg8f/?view_only=b1d12f0f407b457882fcd949f7f59b5c

Pre-registrations

All of our studies were pre-registered. Access pre-registrations through the links below.

Study 1a and 1b	https://aspredicted.org/OSF_IVV
Study 1c	https://aspredicted.org/MKR_VQS
Study 2	https://aspredicted.org/XZW_HKH
Study 3	https://aspredicted.org/4J3_L9N
Study 4	https://aspredicted.org/VKZ_UTW
Study 5	https://aspredicted.org/P6Q_3CR

Appendix B. Study 1

Additional comparisons between apology scenarios

Table B.1 presents summary statistics for all reported feelings across all within-subjects conditions (apology scenarios) for Studies 1a, 1b, and 1c for only participants who recalled a real mutual blame conflict (see next section for same analysis without exclusions).

Table B.1. Study 1 summary statistics for reported feelings in all conditions

Study	Statistic	Condition					
		Self Alone	**Both (Self First)**	**Other Alone**	**Both (Other First)**	**Neither**	**Other Forgives**
1a	M	-18.76	6.98	-5.50	17.67	-15.45	-12.60
	(SD)	(12.01)	(14.78)	(17.12)	(10.68)	(10.13)	(16.15)
1b	M	-2.11	5.08	0.68	7.02	-1.90	2.24
	(SD)	(12.67)	(11.00)	(10.65)	(11.57)	(12.25)	(11.43)
1c	M	-19.68	10.25	-8.88	15.18	-17.18	NA
	(SD)	(11.81)	(14.30)	(15.94)	(11.13)	(11.66)	

We compared the reported feelings for the situation where the participant apologizes alone (Self Alone condition) to the other apology scenarios. As hypothesized in our pre-registration, participants indicated that apologizing first and not getting a return apology would feel, on average, worse than the average feelings associated with apologizing second after the other person apologized first (Both (Other First) condition), 1a: $t(41) = 14.6$, $p < .001$, $d = 2.25$; 1b: $t(155) = 6.8$, $p < .001$, $d = 0.54$; 1c: $t(55) = 14.6$, $p < .001$, $d = 1.94$. As can be seen in Table B.1, participants also reported feeling worse than the average feelings associated with the situation where neither person apologizes (Neither condition). However, these differences were not significant, 1a: $t(41) = 1.7$, $p = .099$, $d = 0.26$; 1b: $t(155) = 0.2$, $p = .835$, $d = 0.02$; 1c: $t(55) = 1.2$, $p = .238$, $d = 0.16$.

We also compared how people would feel when the other person apologizes first and the participant does not return the apology (Other Alone condition), with the case where they do (Both (Other First) condition). As pre-registered, participants reported feeling worse, on average, in the former case than in the latter, 1a: $t(41) = 7.5$, $p < .001$, $d = 1.16$; 1b: $t(155) = 4.8$, $p < .001$, $d = 0.39$; 1c: $t(55) = 10.3$, $p < .001$, $d = 1.38$.

Reported feelings in Studies 1a and 1b including imaginary and one-sided blame conflicts

Table B.2 presents summary statistics for all reported feelings across all within-subjects conditions (apology scenarios) for Studies 1a and 1b, using the larger sample that *includes participants who reported they imagined the conflict scenario and/or the scenario they had in mind involved one-sided blame*. Therefore, in this analysis we only exclude participants who failed the attention check (the only exclusion criterion we pre-registered for these two studies).

Table B.2. Summary statistics for reported feelings in all conditions of Studies 1a and 1b with the sample including people that recalled imaginary or one-sided blame conflict scenarios

Study	Statistic	Condition					
		Self Alone	**Both (Self First)**	**Other Alone**	**Both (Other First)**	**Neither**	**Other Forgives**
1a	M	-18.57	7.55	-5.06	17.49	-14.47	-13.72
	(SD)	(11.89)	(14.28)	(16.69)	(11.14)	(10.95)	(15.89)
1b	M	-3.63	3.48	0.59	7.71	-2.83	1.49
	(SD)	(12.69)	(11.62)	(10.88)	(11.53)	(12.21)	(11.71)

We replicate the main results even when including these observations. As can be seen from Table B.2, participants indicated that apologizing first and not getting a return apology would feel, on average, negative (1a: M = -18.57, SD = 11.89; 1b: M = -3.63, SD = 12.69) and worse than the average positive feelings associated with apologizing first and getting a return apology (1a: M = 7.55, SD = 14.28; 1b: M = 3.48, SD = 11.62;), 1a: $t(46) = 12.9$, $p < 0.001$, $d = 1.88$; 1b: $t(199) = 7.14$, $p < 0.001$, $d = 0.51$. Moreover, people expected that receiving forgiveness in response to their apology would feel worse (1a: M = -13.72, SD = 15.89; 1b: M = 1.49, SD = 11.71) than receiving a return apology, 1a: $t(46) = 8.07$, $p < 0.001$, $d = 1.18$; 1b: $t(199) = 2.23$, $p = 0.027$, $d = 0.16$.

Similar to the previous findings we see that people expect apologizing alone (Self Alone) to feel directionally worse than the status quo when neither person apologizes (Neither scenario). However, these differences are small and either only significant at a 5% level as in Study 1a ($t(46) = 2.14$, $p = .038$, $d = 0.31$), or insignificant and in Study 1b ($t(199) = 0.2$, $p = .927$, $d = 0.07$).

Appendix C. Study 2

Trust game procedure. The first stage of the task was a trust game, which we named the "investment task" for participants. The second mover (trustor) was given 200 chips and asked whether they wanted to keep or pass that amount to the first mover (trustee). If they passed it to their partner, the money would triple to 600 chips, and their partner would choose how to divide the earnings. Second movers completed the stage first and were simply given a binary choice between keeping 200 chips or transferring it to their partner. (We used this because in a pretest most trustors chose to transfer the chips.)

We designed the first mover's version of the stage to encourage them to make unequal splits, returning less to their partners than they kept themselves. First movers were told, "Your match was told that they could give you as many of their chips as they wanted. They were told the number of chips they sent you would be tripled, and then you would decide how to split the chips between yourself and them." They were also told they would learn the amount transferred on the following page, after which they would determine the split. To encourage a selfish split, we borrowed a technique from previous apology research (Leunissen et al., 2012): We created uncertainty about the second mover's initial endowment by telling first movers that their partner had been endowed with somewhere between 200-2500 chips. In pretests, this technique led trustees in a trust game to infer the trustor began with more than 200 chips. Before telling them how much their partner had passed to them, we asked first movers to produce a guess about their partner's initial endowment. First movers then learned that they were transferred 200 chips, which had tripled to 600 chips for them to split.

Participants were presented with seven options for splitting the endowment, three selfish splits (600 for self, 500 for self, 450 for self), one fair split (300 for self), and three generous splits (150 for self, 100 for self, 0 for self). Each choice option displayed how much both parties would earn. However, to further encourage a selfish split, the second mover's amount was calculated after adding in the guess the first mover made about the second mover's initial endowment. For example, if a first mover guessed their partner started out with 800 chips, that would imply their partner kept 600 chips. In that case, the fair split option (300-300) would display the text, "You earn 300; Your match

earns 900." First movers were made aware that their guess about their partner's initial endowment was being used to calculate the final endowments. For screenshots and exact instructions given to participants in the trust game, see the online supplement.

Following the choices in the trust game but before the lying game instructions, we added information to the stage to heighten the possibility that participants would also make the desired choices in the subsequent stage. After making their choice to pass or keep the initial endowment, second movers were told the true fact that, in a pretest of the same task, over 80% of the trustees (the same role their partners were in) kept the majority of the 600 chips for themselves, and that 70% kept 500 or more chips for themselves. This was meant to make second movers feel they had the license to be more selfish in the lying game. After choosing how to split the 600 chips, first movers learned the true initial endowment of the second mover (200 chips) and the true earnings split based on this fact. This information was meant to increase feelings of guilt among first movers in order to make them more inclined to cooperate with the second mover's recommendation in the lying game.

Lying game procedure. The second stage of the task was a lying game, which is a modification of the sender-receiver game (adapted from Gneezy, 2005; see also Crawford & Sobel, 1982). We named this stage the "guessing task" for participants. At the beginning of this stage, both participants were endowed with 400 chips each. In this game, first movers played the "chooser" and had to select between two options, A and B, that would affect how much of their endowed 400 chips each person would get to keep. However, only the second movers, who played the "knower," would have any information about how the options affected their wealth. Second movers would get to send the first mover a recommendation about which option to select. In reality, option A would supply the first mover with 0 chips and the second mover with 400 chips, while option B would supply the first mover with 400 chips and the second mover with 200 chips.

However, to encourage the second mover to recommend the more selfish choice (option A), we introduced uncertainty about option B. While second movers were fully aware of option A's effects, they learned ranges for option B's effects: The first mover would earn anywhere from 0-400 chips, and the second mover would earn anywhere from 0-200 chips. We expected that second movers would feel averse to the risk of option B. After learning the payoffs, the second movers selected

between two messages to send to their partner: “Option A will earn us the most money” and “Option B will earn us the most money.” We expected that the averseness to option B’s risk and a feeling of being slighted in the first task would make second movers more willing to send the misleading message to select option A. First movers were told that their partner knew how the options would affect their earnings and that they would send a recommendation message. After reading the message recommending option A, first movers chose between the two options. For screenshots and exact instructions given to participants in the lying game, see the online supplement. After the lying game, first movers learned the outcomes of both stages. They were reminded of their choice from the trust game, and they learned the truth about the options from the lying game. For screenshots and exact instructions given to participants in the lying game, see the online supplement. For survey files to replicate these tasks, see the .qsf files on the OSF repository.

Figure. C.1. Image shown as manipulation to first movers in Observant condition

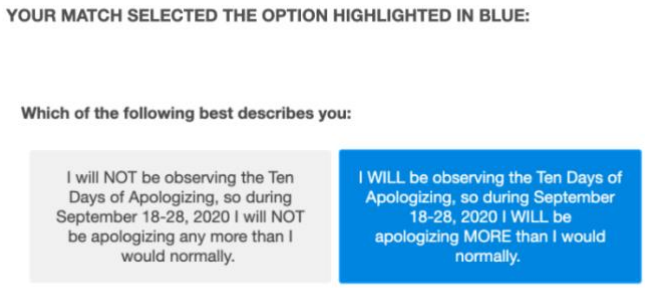
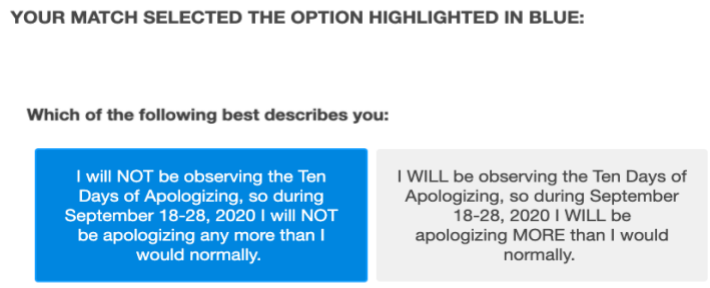


Figure. C.2. Image shown as manipulation to first movers in Nonobservant condition



Main analyses on first mover sample that experienced a mutual blame conflict as defined by choices only (N=288)

After including people who experienced a mutual blame conflict as defined by choices only, the sample was 288 (25% male, M_{age} = 26.84 years, SD_{age} = 9.27). Relative blame was perceived to be similar and approximately 50-50 in both conditions (LOW: M = 49.90, SD = 17.16; HIGH: M = 52.19, SD = 15.82; $t(286) = 1.18$, $p = .24$, $d = 0.14$). Consistent with the intention of the manipulation, participants perceived observant partners to be more likely to send a return apology, M = 2.01, SD = 17.00, than nonobservant partners, M = -2.77, SD = 14.92, $t(286) = 2.53$, $p = .0118$, $d = 0.30$ (using a -30 = "extremely unlikely" to 30 = "extremely likely" sliding scale). There were no cross-condition differences in the overall number of messages sent by first movers (LOW: 68.28%, 99/145; HIGH: 69.93%, 100/143; X^2 (1, N = 288) = 0.09, $p = .7613$, $\Phi = 0.02$; Figure 3). Based on the coding of two RAs blind to our hypothesis, we found that first movers were more likely to apologize when they were partnered with an observant second mover (26.57%, 38/143) than when they were partnered

with a nonobservant second mover (16.55%, 24/145), X^2 (1, N = 288) = 4.28, p = .0386, Φ = 0.12 (Figure 3).

Appendix D. Study 4

Table D.1. Mixed effects regression on likelihood to apologize first

(Intercept)	-22.0491 ***
	(1.7591)
HIGH	21.2987 ***
	(2.2360)
Order (HIGH first)	-0.7013
	(2.5346)
Perceived relative self blame	0.3671 ***
	(0.0496)
HIGH*Order	-3.2165
	(3.2052)
N	300
N (ResponseId)	150
AIC	2488.6449
BIC	2514.5714
R2 (fixed)	0.3943
R2 (total)	0.5099

*** p < 0.001; ** p < 0.01; * p < 0.05.

Preference to apologize first versus second controlling for blame and order of question presentation

We ran an additional mixed-effects regression to assess whether participants were still more likely to apologize second than first after controlling for blame and order in which the apology-likelihood questions were asked. For these, we regressed the likelihood to apologize on a binary variable for the sequence of apology (0 = first, 1 = second). We included a binary variable control for the order in which we asked about the likelihood to apologize first and second (including a main effect and interaction with the binary sequence variable), as well as for perceived relative blame (including an interaction with the binary sequence variable) and participant-level heterogeneity (using random effects; $Pseudo\text{-}R^2_{\text{fixed effects}} = 0.40$).

See Table D.2 for the results. None of the interactions were significant, and we still found that participants reported being less likely to apologize first than second in conflicts they faced in their past where neither person had yet apologized, $t(147.00) = 10.99$, $p < 0.001$, $\beta = 1.16$.

Table D.2. Mixed effects regression on likelihood to apologize

(Intercept)	-15.0131 ***
	(2.1736)
SECOND	26.9744 ***
	(2.4545)
Perceived relative self blame	0.3616 ***
	(0.0745)
Order (SECOND first)	1.1405
	(2.9522)
SECOND*Perceived rel. blame	0.0361
	(0.0842)
SECOND*Order	-3.1555
	(3.3338)
N	300
N (ResponseId)	150
AIC	2569.2979
BIC	2598.9281
R2 (fixed)	0.3993
R2 (total)	0.6170

*** p < 0.001; ** p < 0.01; * p < 0.05.

Appendix E. Study 5

Figure. E.1. Screenshot of message options for first-mover

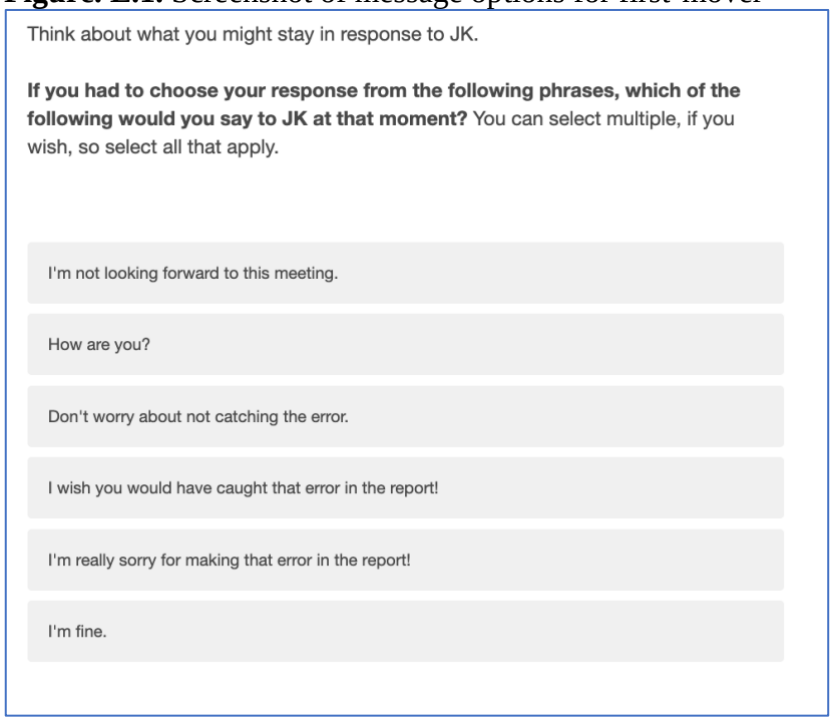


Table E.1. OLS regression on likelihood to apologize

Intercept	0.4068 ***
	(0.0233)
Apologize Second	0.2663 ***
	(0.0330)
Perceived relative self blame	0.8539 ***
	(0.1240)
N	800
R2	0.1333

*** p < 0.001; ** p < 0.01; * p < 0.05.

Wording of message options in Study 5

The apologizing, blaming, and forgiving messages varied slightly by condition to make the phrases more natural in the context: The apology message was either “I'm really sorry for making that error in the report!” (first-mover condition) or “I'm really sorry for actually making that error!” (second-mover condition). The blaming message was either, “I wish you would have caught that error

in the report!" (first-mover condition) or, "Yeah, I wish you would have caught the error!" (second-mover condition). The forgiving message was either, "Don't worry about not catching the error" (first-mover condition) or, "Oh, don't worry about it" (second-mover condition). The three general messages were, "I'm not looking forward to this meeting." "I'm fine," and "How are you?".

Examining the correlation between choice to apologize first and outcome magnitude

We hypothesized that agents' choice to apologize first may not only be sensitive to the probabilities of outcomes but also to the magnitude, i.e., how bad they expect to feel if they end up apologizing alone. We planned to conduct a correlational test of this hypothesis by assessing whether the decision to apologize first among the subset of people who displayed conditional initiation ("conditional initiators") was correlated with the perceived magnitude of the worst outcome (not getting a return apology). To assess perceived outcome magnitude, participants were also asked to indicate how they would generally feel if they were to apologize and not receive a return apology (*feelings question*; sliding scale: -30 = "extremely negative" to 30 = "extremely positive).

We assessed initiator type by asking two binary choice questions with the same prompt ("Which of the following two outcomes would you prefer?") but different options. The options of one question were "Neither [friend's initials] nor I apologize." and "I apologize first, then [friend's initials] apologizes." The options of the other question were "Neither [friend's initials] nor I apologize" and "I apologize first, but [friend's initials] does not apologize." The order of these two questions as well as the order of their options were randomized. Conditional initiators were considered those who chose "I apologize first, then [friend's initials] apologizes," for the first question but "Neither [friend's initials] nor I apologize" for the second.

We ran this test on the subset of participants in the first-mover condition who indicated they were conditional initiators through their response to the two binary choice questions. In a linear regression controlling for perceived relative blame to the self ($R^2 = 0.06$, $F(2, 179) = 6.32$, $p = 0.002$), we did not observe a correlation between the perception of how bad they would feel if they did not get a return apology (feelings question) and the likelihood to apologize first, $\beta = 0.05$, 95% *CI* [-0.08,

0.21]. However, the assumptions for this test were not met—this data suffers from heteroskedasticity. To see this, we plotted the fitted values from this regression against the square root of standardized residuals (scale-location plot) in Figure E.2. We find that the average magnitude of the standardized residuals is increasing as a function of the fitted values (the blue line has a clear upward trend) and that the variability of the standardized residuals' magnitudes varies with the fitted values (the spread of the black dots have a clear hourglass pattern). This is an indication of heteroskedasticity in our data (homoskedasticity would have required a horizontal blue line and a constant spread of the black dots around the blue line). This is because we did not manipulate perceived outcome magnitude which led to a situation where 75% of the ratings fell between -30 and -15, i.e., one fourth of the response scale. Therefore, this should not be taken as evidence that people do not react to the perceived magnitude of the worst outcome. While this underscores that conditional initiators are averse to apologizing alone, it prevents us from testing whether outcome magnitude can influence the decision to apologize first. This suggests that future tests of this hypothesis would be best served by manipulating perceived outcome magnitude to ensure sufficient variance, and also to gather causal evidence.

Figure E.2. Scale-location plot of residuals against the fitted values from regressing the likelihood of apologizing on feelings magnitude (controlling for blame perceptions)