Kobe University Repository : Kernel

PDF issue: 2025-12-04

# Costly apologies communicate conciliatory intention: an fMRI study on forgiveness in response to costly apologies

Ohtsubo, Yohsuke ; Matsunaga, Masahiro ; Tanaka, Hiroki ; Suzuki, Kohta ; Kobayashi, Fumio ; Shibata, Eiji ; Hori, Reiko ; Umemura, Tomohiro ;…

(Citation)
Evolution and Human Behavior,39(2):249-256

(Issue Date)
2018-03

(Resource Type)
journal article

(Version)
Accepted Manuscript

(Rights)
© 2018 Elsevier.
This manuscript version is made available under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/

(URL)
https://hdl.handle.net/20.500.14094/90004772

# Costly Apologies Communicate Conciliatory Intention:

# An fMRI Study on Forgiveness in Response to Costly Apologies

Yohsuke Ohtsubo[1]*, Masahiro Matsunaga[2], Hiroki Tanaka[3], Kohta Suzuki[2], Fumio Kobayashi[2], Eiji Shibata[2], Reiko Hori[2], Tomohiro Umemura[2], and Hideki Ohira[4]

[1] Department of Psychology, Graduate School of Humanities, Kobe University, Nada-ku, Kobe, 657-8501, Japan, [2] Department of Health and Psychosocial Medicine, Aichi Medical University School of Medicine, Nagakute, Aichi, 480-1195, Japan, [3] Brain Science Institute, Tamagawa University, Chiyoda-ku, Tokyo 102-0083, Japan, and [4] Department of Psychology, Graduate School of Environmental Studies, Nagoya University, Chikusa-ku, Nagoya, 464-8601, Japan

* Correspondence should be addressed to Yohsuke Ohtsubo, Graduate School of Humanities, 1-1 Rokkodai-cho, Nada-ku, Kobe, 657-8501, Japan.

Phone: Int+81-78-803-5519; E-mail: yohtsubo@lit.kobe-u.ac.jp

**Acknowledgements**

We are grateful to Haruo Isoda, Akira Ishizuka, and Tamotsu Kawai (Brain and Mind Research Center, Nagoya University, Aichi, Japan) for their technical supports. We also thank Adam Smith for his valuable comments on earlier manuscripts. This work was supported by the Japan Society for the Promotion of Science [KAKENHI 15H03447 to Y.O.]; and the John Templeton Foundation.

***Word count. 6,988 words***

**ABSTRACT**

Reconciliation is an integral part of our social lives. Nevertheless, if a victim perceives the risk of further exploitation by his/her transgressor as non-negligible, the victim may well have difficulty forgiving the transgressor. Therefore, a key ingredient of reconciliation is the transgressor's sincere apology. Theoretical and empirical studies have shown that transgressors can make their apologies credible by incurring a substantial cost. Therefore, we hypothesized that costly apologies, compared to non-costly apologies (i.e., simply saying "sorry"), would effectively communicate a transgressor's conciliatory intention. In a functional magnetic resonance imaging (fMRI) study, participants were asked to imagine a friend committing a mild interpersonal transgression (e.g., standing up the participant) and then apologizing in a costly fashion, apologizing in a non-costly fashion, or not apologizing at all. Compared to non-costly apologies and non-apologies, costly apologies (signals of conciliatory intention) more strongly activated the theory-of-mind network (i.e., bilateral temporoparietal junction, precuneus, medial prefrontal cortex). Moreover, we did not observe any significant differences in brain responses to non-costly apologies and non-apology controls. These results underscore the importance of costly signals in human communication and in human peace-making in particular.

**Key words**: costly apology, costly signaling, reconciliation, forgiveness, theory of mind

## 1. Introduction

There is growing interest in the study of reconciliation and forgiveness in the evolutionary behavioral sciences (e.g., Aureli & de Waal, 2000; McCullough, 2008; McCullough et al., 2012; Silk, 2002). On the one hand, maintaining reliable relationships with social partners is an adaptive strategy in most primate societies. On the other hand, conflicts over relatively minor resources are prevalent and unavoidable. Accordingly, it has been hypothesized that social primates are more prone to reconciling with valuable partners than with less valuable partners (de Waal, 2000). This so-called valuable relationships hypothesis has been validated in research on multiple primate species, such as long-tailed macaques (Cords & Thurnheer, 1993) and chimpanzees (Watts, 2006). Recently, McCullough et al. (2010) applied this hypothesis to humans and uncovered that people are also more likely to forgive valuable, as opposed to less valuable, partners. Furthermore, Burnette et al. (2012) found that the effect of relationship value was substantially diminished when people were afraid that they might be exploited by the same partner again. Therefore, if transgressors wish for forgiveness from their victims, they must not only become a valuable partners but also somehow reduce the degree to which their victims perceive them as an exploitation risk.

### *1.1. Costly apology as an honest signal of conciliatory intention*

In a now-classic essay, Goffman (1971) maintained that apologies facilitate interpersonal reconciliation. Decades of subsequent empirical studies have confirmed this argument (e.g., Darby & Schlenker, 1982; Fischbacher & Utikal, 2013; McCullough et al., 1997; Ohbuchi et al., 1989). According to a recent meta-analytic review (Fehr et al., 2010), the effect size of an apology on forgiveness, indicated by a correlation coefficient ($r$), is estimated to be as high as .42. However, simply making apologetic remarks (e.g., “I’m sorry,” “I’ll never do it again”) is sometimes insufficient to convey a transgressor’s sincere intention, thus failing to lead to

forgiveness from the victim (e.g., Risen & Gilovich, 2007; Skarlicki et al., 2004). This is because talk is cheap (Bottom et al., 2002). In particular, it is as easy for malevolent transgressors, who intend to exploit their victim again, to make such remarks as it is for benign transgressors, who sincerely wish to restore the relationship.

Game theoretic analyses of honest signaling suggest that apologetic signals can be made credible if the transgressor incurs a sufficient cost in producing them (Ho, 2012; Martinez-Vaquero et al., 2015; Ohtsubo & Watanabe, 2009). Suppose that a transgressor obtained a benefit of $b_e$ from a single instance of exploitation. If the transgressor forfeits $b_e$ to make an apology, his/her lack of exploitative intention is transparent. Moreover, costly apologies implicitly reveal the transgressor's valuation of the relationship. Therefore, transgressors have an incentive to make costly apologies if and only if they expect to receive a long-term benefit from the relationship that is greater than the cost of the apology. This model of costly apology leads to what we term the *conciliatory intention signaling hypothesis*: A sufficiently costly apology serves as an honest signal of a transgressor's sincere intention to restore the endangered relationship (or as an honest signal of non-exploitative intention).

The conciliatory intention signaling hypothesis, which was derived from signaling theory (Searcy & Nowicki, 2005; Zahavi & Zahavi, 1997), assumes the co-evolution of signal senders' propensity to produce a costly signal under relevant circumstances and signal receivers' sensitivity to a particular type of signal. In the apology-making context, the signal senders and receivers correspond to transgressors and victims, respectively. Thus, the conciliatory intention signaling hypothesis predicts that (i) transgressors (signal senders) are more likely to apologize in a costly manner when they value the victim than when they do not, and that (ii) victims (signal receivers) perceive costly apologies as more sincere than non-costly apologies. Empirical evidence corroborates these predictions of the conciliatory intention signaling hypothesis.

First, after causing harm to valuable partners, signal senders (transgressors) tend to feel guilty (Nelissen, 2014) and become increasingly willing to make costly apologies (Ohtsubo & Yagi, 2015). In a vignette study, for example, participants were asked to imagine that they had committed a mild transgression against one of their real friends. Participants' valuation of the friend was separately measured along with their willingness to make various forms of costly apologies. In line with the conciliatory intention signaling hypothesis, the relationship value of the friend was positively correlated with the participant's willingness to make costly apologies (Ohtsubo & Yagi, 2015, Studies 1 and 2). Furthermore, an autobiographical recall study confirmed that participants had made costly apologies more frequently after causing harm to valuable, as opposed to less valuable, partners (Ohtsubo & Yagi, 2015, Study 3).

Second, after being harmed, signal receivers (victims) are sensitive to the costliness of their transgressor's conciliatory behaviors. Consider compensation, a common costly conciliatory behavior. Although it is usually an effective means of eliciting forgiveness, victims are less forgiving when they receive insufficient compensation (relatively low-cost compensation) than when they receive sufficient compensation (relatively high-cost compensation) (Desmet et al., 2010, 2011). More importantly, it appears that victims are not solely concerned with recovering access to lost benefits; they are also concerned with receiving costly signals of conciliatory intention, such as the cancelation of important business so as to immediately apologize (Ohtsubo & Watanabe, 2009), and the infliction of financial or physical self-punishment (Inbar et al., 2013; Nelissen & Zeelengerg, 2009; Watanabe & Ohtsubo, 2012). Conducting a series of vignette experiments and an economic game experiment, Ohtsubo and Watanabe (2009) showed that victims perceive such unilaterally costly apologies as well as costly compensation, as being more sincere than non-costly apologies, and become more forgiving after receiving such costly conciliatory signals. Cross-cultural research corroborated this effect in Chile, China, Indonesia,

Japan, the Netherlands, South Korea, and the United States (Ohtsubo et al., 2012).

*1.2. Theory-of-mind network and social/communicative intention*

The aforementioned lines of research provide converging support for the conciliatory intention signaling hypothesis. Nevertheless, in previous studies, signal receivers' reactions to costly apologies were assessed by standard social psychological measures, such as self-report forgiveness and behaviors in economic games. In the animal signaling literature, signals are defined as "any act or structure which alters the behaviour of other organisms, which evolved because of that effect, and which is effective because the receiver's response has also evolved" (Maynard Smith & Harper, 2003, p. 3). This definition, especially the last part, necessitates stringent tests on signal receivers' responsiveness to signals. If transgressors' propensity to make costly apologies and victims' responsiveness in fact co-evolved like other animal signaling systems, victims' responsiveness to costly apologies must have some biological basis. In particular, we predicted that some regions of the brain would more strongly respond to costly apologies than to non-costly apologies.

A plausible candidate region that we expected to respond to costly apologies was the theory-of-mind (ToM) network. Conciliatory intention communicated by costly apologies is conceivable as being both social and communicative (Ciaramidaro et al., 2007). Unlike a private intention (e.g., to have a cup of coffee), a conciliatory intention (e.g., to restore a relationship with a victim) is social because it cannot be accomplished in isolation: no matter how strongly a transgressor intends to reconcile with his/her victim, he/she cannot achieve this goal unless the victim reacts to his/her apology favorably. Conciliatory intention is communicative as well because costly apologies deliver information to the signal receiver about the signal sender's internal state (a sincere intention to restore the relationship). Ciaramidaro et al. (2007) showed that a social and communicative intention recruits the bilateral temporoparietal junction (TPJ),

the precuneus, and the medial prefrontal cortex (MPFC). Therefore, we predicted that costly apologies would activate these four ToM-related regions more strongly than non-costly apologies would.

No previous neuroscientific research investigated forgiveness processes from the perspective of signaling theory. Most neuroscientific research into forgiveness has investigated brain activity associated with responses to norm violations and the degree to which such violations are forgivable from a third-person perspective (Farrow et al., 2001, 2005; Hayashi et al., 2010; Young & Saxe, 2009). Moreover, when forgiveness was investigated in the context of interpersonal relationships (Ricciardi et al., 2013), the researchers did not particularly focus on apologies (see Billingsley & Losin, 2017, for a review). One exceptional study, however, did investigate the neural basis of forgiveness after receiving an apology (Strang et al., 2014). This study revealed that forgiveness was positively correlated with activation in a sub-region of the right TPJ. Therefore, Strang et al.'s result conceivably offers supportive evidence that apologies convey conciliatory intention. However, they did not manipulate the costliness of apologies in their study. In the present fMRI study, we examined differences in brain responses to scenarios describing costly and non-costly apologies.

*1.3. Alternative hypotheses*

However, we recognized that there were at least two alternative hypotheses. First, costly signals are often interpreted as indicators of a signaler's quality/traits (Barclay, 2016; Gangestad & Thornhill, 2007), but not intention. In the context of apology-making, costly apologies might communicate non-exploitative personality traits, such as agreeableness (Tabak et al., 2012). Neuroscientific research has indicated that trait attributions recruit the MPFC, but not the TPJ (Saxe, 2006; Van Overwalle, 2009). Therefore, if costly apologies merely communicate the signaler's quality/traits, we expected that the TPJ would not be activated. According to this first

alternative hypothesis, the signal receiver's (i.e., victim's) psychology is equipped with a domain-specific trait inference mechanism, instead of the hypothesized intention-reading mechanism. However, we prioritized the conciliatory intention signaling hypothesis because systematic research on attribution has revealed that upon observing others' behaviors, people engage in mental state inferences first and trait inferences later (Malle & Holbrook, 2012).

The second alternative hypothesis is that non-costly, verbal apologies are sufficient to communicate conciliatory intention. This implies that humans do not have domain-specific signals in the context of reconciliation. Rather, humans rely on a domain-general communicative tool (i.e., language) to facilitate the reconciliation process. For this to be valid, there would be no differences in brain responses (in the ToM network in particular) to scenarios describing costly and non-costly apologies. This second alternative hypothesis seems reasonable because humans are a profoundly linguistic species (Pinker, 1994). Nevertheless, we prioritized the conciliatory intention signaling hypothesis because, as we explained, previous studies have consistently shown that costly apologies are more effective in communicating conciliatory intention than non-costly, verbal, apologies are (Ohtsubo & Watanabe, 2009; Ohtsubo et al., 2012).

**2. Pilot study**

The functional magnetic resonance imaging (fMRI) study consisted of 30 hypothetical forgiveness judgments (see Appendix for the scenarios). Participants were asked to think of a specific same-sex friend and assume that this friend committed a series of hypothetical transgressions. For each judgment, participants were presented a hypothetical transgression (e.g., your friend stood you up) and the friend's reaction, which was either a costly apology (e.g., your friend apologized and treated you to lunch), a non-costly apology (e.g., your friend apologized, saying he/she was just kidding), or a non-apology (e.g., your friend did not apologize, saying it was no big deal). Although we wrote the scenarios by referring to previous studies (Ohtsubo &

Watanabe, 2009; Ohtsubo et al., 2012), these scenarios were much shorter than those in the original studies (see Supplementary Materials). Therefore, to confirm that the brief scenarios retained the effect observed in previous studies, we conducted an online pilot study involving 300 respondents (150 females and 150 males, mean ± *SD* = 20.4 ± 1.17 years).

Each respondent read three transgression scenarios (of a total of 10 scenarios). Each transgression scenario was followed by either a costly apology, a non-costly apology, or a non-apology reaction. The order of the three types of transgressor reactions was randomized (see Supplementary Materials for more details). To assess whether the costly apology scenarios contained sufficient information for respondents to make mental state inferences, we had them rate the sincerity of the reaction and perceived risk of further exploitation. Mean sincerity (± *SD*) was 3.44 ± 1.04, 3.25 ± 1.05, and 2.51 ± 0.98 in the costly, non-costly, and non-apology conditions, respectively ($F_{2, 598} = 92.12$, $p < .001$, $\eta_p^2 = 0.24$). As expected, mean sincerity in the costly apology condition was significantly greater than mean sincerity in the non-costly apology condition ($t_{598} = 2.71$, $p = .007$) and the non-apology condition ($t_{598} = 10.16$, $p < .001$). Mean perceived exploitation risk was 2.86 ± 1.04, 3.07 ± 1.08, and 3.32 ± 1.08 in the costly, non-costly, and non-apology conditions, respectively ($F_{2, 598} = 15.67$, $p < .001$, $\eta_p^2 = 0.06$). Again, confirming our prediction, perceived exploitation risk was lower in the costly apology condition than in the non-costly apology condition ($t_{598} = 2.96$, $p = .003$) and non-apology condition ($t_{598} = 3.48$, $p < .001$). Hence, it can be said that the costly apology scenarios contained sufficient information that allowed respondents to make mental state inferences. However, the critical question for the signaling hypothesis is whether people spontaneously engage in mental state inferences in response to costly apologies without being explicitly instructed to do so. We conducted an fMRI experiment to answer this question.

**3. Methods**

*3.1. Participants*

Forty healthy volunteers (30 females and 10 males, 20.3 ± 1.52 years) participated in the study. Two female participants were discarded due to left-handedness, and one additional female participant was discarded due to excessive head movement (more than 3 mm) during the fMRI scan. As a result, 37 participants were retained for the data analysis. Participants provided informed consent, and then filled out a set of questionnaires, including a Japanese version of the trait forgivingness scale (J-TFS) (Berry et al., 2005; Ohtsubo et al., 2015).

*3.2. Task*

In the imaging study, participants were first presented one of 10 transgression scenarios (e.g., [Your friend] forgot that the two of you promised to meet, and ended up standing you up), followed by an apology (or non-apology) scenario (see Fig. 1). Participants then indicated their willingness to forgive their friend using a Visual Analogue Scale (VAS) slider, which was controlled by a set of response buttons (pressing the left/right button shifted the slider in the left/right direction, respectively). The two poles of the slider were denoted as “not at all forgivable” (converted to 0) and “completely forgivable” (converted to 100). This procedure was repeated 30 times. Participants were presented 10 costly apology scenarios, 10 non-costly apology scenarios, and 10 non-apology scenarios, which were treated as a within-participant factor. Three functional imaging runs (each consisting of 10 trials and lasting about 6 minutes) were performed for each participant. The order of the 30 scenarios was pre-randomized, with participants receiving either a forward (scenarios 1–30) or reversed (scenarios 30–1) version so as to mitigate possible order effects.

*3.3. fMRI data acquisition*

Functional neuroimaging was conducted using a 3-Tesla MRI scanner (Verio; Siemens Ltd., Erlangen, Germany) at the Brain and Mind Research Center, Nagoya University, Japan. Each

participant's head was immobilized within a 32-element phased-array head coil. Imaging was performed using an echo-planar imaging (EPI) gradient-echo sequence (echo time [TE] = 30 ms, repetition time [TR] = 2500 ms, field of view [FOV] = 192 × 192 $mm^2$, flip angle = 80º, matrix size = 64 × 64, 39 slices, slice thickness = 3 mm, total number of volumes = 148). A whole-brain, high-resolution T1-weighted anatomical magnetization-prepared rapid-acquisition gradient echo (MP-RAGE) MRI was also acquired for each participant (TE = 1.98 ms, TR = 1800 ms, FOV = 256 × 256 $mm^2$, flip angle = 9º, matrix size = 256 × 256 pixels, and slice thickness = 1 mm).

*3.4. fMRI data preprocessing and analyses*

We used Statistical Parametric Mapping (SPM) software (SPM12 revision 6225; The Wellcome Department of Cognitive Neurology, London, UK) implemented in MATLAB 2014b (MathWorks Inc., Massachusetts) to analyze the functional images. The first four volumes of each fMRI run were discarded due to unsteady magnetization. After all of the volumes were realigned, differences in slice timing within each image volume were corrected. The reference image was the center of the volume. The whole-head 3D MP-RAGE volume was co-registered with the EPI volumes, and the whole-head 3D MP-RAGE volume was normalized in accordance with the Montréal Neurological Institute (MNI) T1 image template (ICBM152) using a non-linear basis function. Subsequently, normalization parameters were applied to all of the EPI volumes. The normalized EPI images were then spatially smoothed in three dimensions using an 8-mm full-width at half-maximum Gaussian kernel. After the realignment processes, we checked for head movements greater than 3 mm during the experimental run. Task-related activation was statistically evaluated on a voxel-by-voxel basis using the general linear model at the individual level to generate contrast images. The transgression (10 s), friend's apology or non-apology (10 s), and rating phases (5 s) were separately modeled by a block design convolved with a canonical hemodynamic response. The transgression and rating phases were included as covariates of no

interest in order to partial out their contribution to brain activation in single participant analyses. Using contrast images relating to the reaction phase of the three conditions (Costly Apology, Non-costly Apology, Non-apology), we conducted a random-effects analysis at the group level, with a within-subject design ANOVA. The statistical threshold was set at an uncorrected $p < .001$ at the voxel level and a familywise-error (FWE) corrected $p < .05$ at the cluster level (whole brain).

*3.5 Ethical statement and data availability*

This study was approved by the ethical review board at the first author's institute. The data sets associated with the pilot study and the main study are available as Supplementary Data.

**4. Results**

*4.1. Self-reported forgiveness*

We first analyzed self-reported forgiveness scores. As shown in Fig. 2A, participants' willingness to forgive was highest in response to costly apologies ($M \pm SD = 70.15 \pm 13.80$), middle for non-costly apologies (51.19 ± 14.43), and lowest for non-apologies (25.06 ± 8.29), $F_{2, 72} = 269.12$, $p < .001$, $\eta_p^2 = 0.88$. Importantly, forgiveness was significantly higher in the costly apology condition than in the non-costly apology ($t_{72} = 13.39$, $p < .001$) and non-apology ($t_{72} = 23.10$, $p < .001$) conditions. In addition, we assayed each participant's trait forgivingness (i.e., general tendency to forgive others) using the J-TFS. The 10 items in the J-TFS were aggregated to obtain the single score of trait forgivingness (Cronbach's $\alpha = .68$). We found that participants high in trait forgivingness were more likely to forgive their transgressor in all three conditions (Figs. 2B to 2D).

*4.2. fMRI data*

The fMRI data were analyzed using the subtraction method to test whether costly apologies uniquely recruited activity in the focal regions underlying ToM. Accordingly, we conducted the

following three comparisons: (Costly Apology > Non-costly Apology), (Costly Apology > Non-apology), and (Non-costly Apology > Non-apology). The statistical threshold was set at an uncorrected $p < .001$ at the voxel level and a FWE corrected $p < .05$ at the cluster level (whole brain). As predicted, in the former two comparisons, we observed significant increases in blood-oxygen-level dependent (BOLD) responses in the bilateral TPJ, precuneus, and MPFC (Fig. 3, Table 1). In other words, the four characteristic regions that respond to social and communicative intention were more strongly recruited in the costly apology condition than in the non-costly apology and non-apology conditions.

Intriguingly, there were no significant differences in any part of the brain in the third comparison; that is, the brain did not appear to distinguish between non-costly apologies (the transgressor saying “sorry”) and non-apologies (the transgressor doing something other than saying “sorry”). Furthermore, as Table 1 shows, some areas other than the predicted ToM network (e.g., the dorsolateral prefrontal cortex, orbitofrontal cortex, cerebellum) were significantly more active in the costly apology condition than in the other two conditions. We will address these issues in the Discussion section.

**5. Discussion**

Our study tested the conciliatory intention signaling hypothesis as it applies to costly apologies. Specifically, we investigated whether costly apologies, the presumed function of which is to communicate a sincere conciliatory intention, recruit regions of the ToM network implicated in research on social/communicative intention (bilateral TPJ, precuneus, and MPFC). Our results support this prediction. Although participants were not explicitly asked to infer the transgressor’s intention (they were merely asked to imagine the described situations and decide how willing they would be to forgive the transgressor), the ToM network was significantly more active in response to costly apologies than non-costly apologies and non-apologies. The engagement of the

bilateral TPJ contradicts the first alternative hypothesis that costly apologies communicate the signaler's personality traits (e.g., agreeableness), but not their intention. Likewise, significant differences in ToM network responses between the costly apology and non-costly apology conditions contradict the second alternative hypothesis, and this may be surprising for researchers who expect that linguistic signals are sufficient to communicate sincere intention. This implies that the various types of costs involved in our study's scenarios (time, money, effort), but not apologetic statements alone, are crucial to the activation of the ToM network. This result is consistent with the idea that a domain-specific costly signal (i.e., a costly form of apologies) co-evolved with the signal receiver's responsiveness.

It is noteworthy that we did not find any significant differences between the non-costly apology and non-apology conditions in terms of brain activity. This is somewhat perplexing because participants reported higher levels of forgiveness in response to non-costly apologies than to non-apologies. One possible explanation is that the two conditions were practically equivalent. Although no explicit statements of apology were included in the non-apology scenarios, most of them included some excuses (e.g., "it was just a joke," "[I] simply forgot to invite you"). In Schönbach's (1990) research on accounts, both apologies and excuses were considered as more mitigating accounts than justifications and denials. Therefore, in future studies, more appropriate control scenarios, such as scenarios describing justifications and denials, or some other equally complex but non-social situations, must be included.

Some readers might be perplexed by the main finding—costly apologies more strongly activated the ToM network than non-costly apologies and non-apologies did. It is plausible that non-apologies and non-costly apologies also invoke mental state inferences. If your transgressor does not apologize, you would probably think about whether he/she intentionally committed the transgression. If your transgressor verbally apologizes in a non-costly manner (e.g., with a simple

"sorry"), you would probably question his/her sincerity. Given these putative reactions to non-apologies and non-costly apologies, one question arising naturally is why they failed to activate the ToM network in the present fMRI study. A possible explanation relates to the scenario-based methodology; that is, participants did not face a real exploitation risk. Therefore, they might not have engaged in effortful reasoning processes. Notice, however, that this explanation does not discredit the importance of the reported result. If participants avoided effortful mind-reading in response to non-apologies and non-costly apologies, we need further explanation for why participants did *not* avoid it in response to costly apologies. If signal receivers possess some evolved (presumably "automatic") responsiveness, this explains the observed difference between the costly apology conditions and the other two conditions.

The present results illuminate the importance of signaling theory in understanding both human reconciliation and mind-reading processes. Despite possessing a highly evolved linguistic ability, humans seem to still depend on an evolutionarily ancient mode of communication—costly signals (Searcy & Nowicki, 2005; Zahavi & Zahavi, 1997). Conciliatory signals require that a signaler commits to a peaceful course of interaction with his/her former victim. However, because talk is cheap, merely saying "sorry" can fail to convey a sincere intention to restore an endangered relationship. In such cases, conciliatory signals must be costly. The cost of an apology makes transparent the transgressor's benign, non-exploitative intent, which the victim cannot directly observe. This understanding also provides a novel insight for the mind-reading literature. Although typical ToM research conceptualizes mind-reading as a solitary cognitive activity of the mind-reader (Malle, 2004; Epley & Waytz, 2010), the signaling theory suggests that mind-reading should be viewed as a cooperative activity involving both a mind-reader and a person being read (cf. Schilbach, 2015).

In the Results section, we noted that in addition to the hypothesized ToM network, there

were other regions of the brain that selectively responded to costly apologies (see Table 1). The orbitofrontal cortex (OFC) was among such unpredicted regions. Although OFC relates to various functions (Stalnaker et al., 2015), one of its functions is to encode economic value or utility (e.g., McNamee et al., 2013; Padoa-Schioppa & Assa, 2006; Plassmann et al., 2007). Interestingly, Tooby et al. (2008) argued that various forms of interpersonal interactions convey information regarding relationship value (e.g., whether this partner is likely to deliver benefits to me), and humans are equipped with some psychological mechanism to update the relationship value (or "welfare tradeoff ratio" in their terminology) of a specific partner. Consistent with this thesis, Forster and McCullough (2017) observed that victims upregulated transgressors' relationship value in response to costly apologies from the transgressors. Therefore, it is possible that the OFC activation reflects the spontaneous recalibration of the transgressor's utility. This idea needs to be more closely scrutinized using stimuli exclusively associated with relationship value.

In sum, the present study confirmed the conciliatory intention signaling hypothesis. As predicted, costly apologies engaged brain regions that are crucial to ToM (i.e., bilateral TPJ, precuneus, and MPFC). Although saying "I'm sorry" fosters forgiveness to some extent, *costly* actions seem to speak louder than words in the reconciliation process.

**Appendix**

The following scenarios are listed in order of "transgression," "costly apology," "non-costly apology," and "non-apology." Although they noticeably vary in word count, variance in the Japanese originals was kept to a minimum (see Table S1 in Supplementary Data).

1. (Your friend) forgot that the two of you promised to meet, and ended up standing you up.

   (Your friend) apologized and treated you to lunch.

(Your friend) apologized for forgetting about the promise.

(Your friend) did not apologize, simply saying he/she forgot about the meeting.

2. (Your friend) told some other friends about someone you had a crush on.

(Your friend) apologized for his/her carelessness, and made sure that everyone he/she told about your crush, would keep it secret.

(Your friend) apologized, saying he/she had been careless.

(Your friend) did not apologize, saying it’s not something you need to worry about.

3. (Your friend) borrowed something from you, and then lent it to someone else without your permission.

(Your friend) apologized, and immediately recovered your possession from the person he/she lent it to.

(Your friend) apologized, saying he/she felt simply obliged to let the other person borrow your possession.

(Your friend) did not apologize, saying it’s no big deal because the other person is also your friend.

4. (Your friend) made fun of your appearance while drinking at a party.

(Your friend) apologized, and told you that he/she would give up drinking even though he/she really liked to drink.

(Your friend) apologized, saying he/she didn't mean to offend you.

(Your friend) said not to worry, that it was just a joke.

5. (Your friend) created a LINE* group for your mutual friends, but did not invite you.

(Your friend) came to you to make an apology as soon as he/she noticed that you had not been invited.

(Your friend) apologized, saying he/she simply forgot to invite you.

(Your friend) said he/she simply forgot to invite you, that it wasn't on purpose.

* LINE is one of most popular social networking services (e.g. Facebook) in Japan.

6. (Your friend) borrowed something from you, but treated it poorly and ended up breaking it.

(Your friend) bought a replacement for your broken possession, and gave it to along with an apology.

(Your friend) apologized, saying he/she did not mean to break it.

(Your friend) returned the object to you, saying, "looks like it's broken."

7. (Your friend) did not invite you when he/she made plans with your mutual friends.

(Your friend) planned another get-together, as a means to apologize for his/her oversight.

(Your friend) apologized, saying he/she thought you were busy.

(Your friend) did not apologize, saying he/she thought you wouldn't have been interested.

8. (Your friend) posted a dirty joke using your twitter account.*

(Your friend) apologized, and said he/she would do anything to make it up to you.

(Your friend) apologized, saying he/she was just kidding around.

(Your friend) did not apologize, saying anyone would know it's just a joke.

* In the Japanese scenario, it was phrased as 'posted a dirty joke from your smartphone'. In Japanese, this implies that the friend posted from the participant's account.

9. (Your friend) told some people about a problem that you were trying to keep secret.

(Your friend) apologized, and spent the entire night helping you find a solution to the problem.

(Your friend) apologized, saying he/she thought you wouldn't mind other people knowing.

(Your friend) did not apologize, saying it's no big deal.

10. (Your friend) was supposed to meet you somewhere, but arrived an hour late.

(Your friend) apologized, for being late, and treated you to something to eat.

(Your friend) apologized, saying he/she simply overslept.

(Your friend) simply said he/she overslept, and did not apologize.

**References**

Aureli, F. & de Waal, F. B. M. (eds.). (2000). *Natural conflict resolution*. Berkley, CA: University of California Press.

Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology*, *7*, 33-38.

Berry, J. W., Worthington, E. L. Jr., O'Connor, L. E., Parrott, L., III, & Wade, N. G. (2005). Forgivingness, vengeful rumination, and affective traits. *Journal of Personality*, *73*, 183-225.

Billingsley, J., & Losin, E. A. R. (2017). The neural systems of forgiveness: An evolutionary psychological perspective. *Frontiers in Psychology*, *8*:737.

Bottom, W. P., Gibson, K., Daniels, S. E., & Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organization Science*, *13*, 497-513.

Burnette, J. L., McCullough, M. E., Van Tongeren, D. R., & Davis, D. E. (2012). Forgiveness results from integrating information about relationship value and exploitation risk. *Personality and Social Psychology Bulletin*, *38*, 345-356.

Ciaramidaro, A. Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B. G., et al. (2007). The intentional network: How the brain reads varieties of intentions. *Neuropsychologia*, *45*, 3105-3113.

Cords, M., & Thurnheer, S. (1993). Reconciling with valuable partners by long-tailed macaques. *Ethology*, *93*, 315-325.

Darby, B. W., & Schlenker, B. R. (1982). Children's reactions to apologies. *Journal of Personality and Social Psychology*, *43*, 742-753.

Desmet, P. T. M., De Cremer, D., & van Dijk, E. (2010). On the psychology of financial

compensation to restore fairness transgression: When intentions determine value. *Journal of Business Ethics*, *95*, 105-115.

Desmet, P. T. M., De Cremer, D., & van Dijk, E. (2011). In money we trust? The use of financial compensations to repair trust in the aftermath of distributive harm. *Organizational Behavior and Human Decision Processes*, *114*, 75-86.

de Waal, F. B. M. (2000). Primates--A natural heritage of conflict resolution. *Science*, *289*, 586-90.

Epley, N., & Waytz, A. (2010). Mind perception. In: S. F. Fiske, D. T. Gilbert, G. Lindzey, (Eds.). *Handbook of social psychology* (5th ed., vol. 1) (pp. 4998-541). Hoboken, NJ: Wiley.

Farrow , T. F. D., Hunter, M. D., Wilkinson, I. D., Gouneea, C., Fawbert, D., Smith, R., et al. (2005). Quantifiable change in functional brain response to empathic and forgivability judgments with resolution of posttraumatic stress disorder. *Psychiatry Research: Neuroimaging*, *140*, 45-53.

Farrow, T. F. D., Zheng, Y., Wilkinson, I. D., Spence, S. A., Deakin, J. F. W., Tarrier, N., et al. (2001). Investigating the functional anatomy of empathy and forgiveness. *Neuroreport*, *12*, 2433-2438.

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

Fischbacher, U., & Utikal, V. (2013). On the acceptance of apologies. *Games and Economic Behavior*, *82*, 592-608.

Forster, D., & McCullough, M. E. (2017). *Do people's perceptions of a transgressor's relationship value cause forgiveness?* Paper presented at the 29th annual conference of the Human Behavior and Evolution Society, Boise, ID.

Gangestad, S. W., & Thornhill, R. (2007). The evolution of social inference processes: The

importance of signaling theory. In: J. P. Forgas, M. G. Haselton, W. von Hippel (Eds). *Evolution and the social mind: Evolutionary psychology and social condition* (pp. 33-48). New York: Psychology Press.

Goffman, E. (1971). *Relations in public: Microstudies of the public order*. New York: Harper & Row.

Hayashi, A., Abe, N., Ueno, A., Shigemune, Y., Mori, E., Tashiro, M., et al. (2010). Neural correlates of forgiveness for moral transgressions involving deception. *Brain Researc*h, *1332*, 90-99.

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

Inbar, Y., Pizarro, D. A., Gilovich, T., & Ariely, D. (2013). Moral masochism: On the connection between guilt and self-punishment. *Emotion*, *13*, 14-18.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, *102*, 661- 684.

Martinez-Vaquero, L. A., Han, T. A., Pereira, L. M., & Lenaerts, T. (2015). Apology and forgiveness evolve to resolve failures in cooperative agreements. *Scientific Reports*, *5*: 10639.

Maynard Smith, J., & Harper, D. (2003). *Animal signals*. Oxford, UK: Oxford University Press.

McCullough, M. E. (2008). *Beyond revenge: The Evolution of the forgiveness instinct*. San Francisco, CA: Jossey-Bass.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2012). Cognitive systems for revenge and

forgiveness. *Behavioral and Brain Sciences*, *36*, 1-58.

McCullough, M. E., Luna, L. R., Berry, J. W., Tabak, B. A., & Bono, G. (2010). On the form and function of forgiving: Modeling the time-forgiveness relationship and testing the valuable relationships hypothesis. *Emotion*, *10*, 358-376.

McCullough, M. E., Worthington, E. L., Jr., & Rachal, K. C. (1997). Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology*, *73*, 321-336.

McNamee, D., Rangel, A., & O'Doherty, J. P. (2013). Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nature Neuroscience*, *16*, 479-485.

Nelissen, R. M. A. (2014). Relational utility as a moderator of guilt in social interaction. *Journal of Personality and Social Psychology*, *106*, 257-271.

Nelissen, R. M. A., & Zeelenberg, M. (2009). When guilt evokes self-punishment: Evidence for the existence of a Dobby effect. *Emotion*, *9*, 118-122.

Ohbuchi, K., Kameda, M., & Agarie, N. (1989). Apology as aggression control: Its role in mediating appraisal of and response to harm. *Journal of Personality and Social Psychology*, *56*, 219-227.

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior*, *30*, 114-123.

Ohtsubo, Y., Watanabe, E., Kim, J., Kulas, J. T., Muluk, H., Nazar, G., et al. (2012). Are costly apologies universally perceived as being sincere? A test of the costly apology-perceived sincerity relationship in seven countries. *Journal of Evolutionary Psychology*, *10*, 187-204.

Ohtsubo, Y., & Yagi, A. (2015). Relationship value promotes costly apology-making: Testing the valuable relationships hypothesis from the perpetrator's perspective. *Evolution and Human Behavior*, *36*, 232-239.

Ohtsubo, Y. Yamaura, K., & Yagi, A. (2015). Development of Japanese measures of reconciliatory tendencies: The Japanese Trait Forgivingness Scale and the Japanese Proclivity to Apologize Measure. *Japanese Journal of Social Psychology, 31*, 135-142.

Padoa-Schioppa, C., & Assa, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature, 441*, 223-226.

Pinker, S. (1994). *The language instinct: How the mind creates language*. New York: William Morrow.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience, 27*, 9984-9988.

Ricciardi, E., Rota, G., Sani, L., Gentili, C., Gaglianese, A. Guazzelli, M., & Pietrini, P. (2013). How the brain heals emotional wounds: The functional neuroanatomy of forgiveness. *Frontiers in Human Neuroscience, 7*: 839.

Risen, J. L., & Gilovich, T. (2007). Target and observer differences in the acceptance of questionable apologies. *Journal of Personality and Social Psychology, 92*, 418-333.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology, 16*, 235-239.

Schilbach, L. (2015). Eye to eye, face to face and brain to brain: Novel approaches to study the behavioral dynamics and neural mechanisms of social interactions. *Current Opinion in Behavioral Sciences, 3*, 130-135.

Schönbach, P. (1990). *Account episodes: The management or escalation of conflict*. Cambridge, UK: Cambridge University Press.

Searcy, W. A., & Nowicki, S. (2005). *The evolution of animal communication: Reliability and deception in signaling systems*. Princeton, NJ: Princeton University Press.

Silk, J. B. (2002). The form and function of reconciliation in primates. *Annual Review of*

*Anthropology*, *31*, 21-44.

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*, 322-341.

Stalnaker, T. A., Cooch, N. K., & Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. *Nature Neuroscience*, *18*, 620-627.

Strang, S., Utikal, V., Fischbacher, U., Weber, B., & Falk, A. (2014). Neural correlates of receiving an apology and active forgiveness: An fMRI study. *PLoS ONE*, *9*(2): e87654.

Tabak, B. A., McCullough, M. E., Luna, L. R., Bono, G., & Berry, J. W. (2012). Conciliatory gestures facilitate forgiveness and feelings of friendship by making transgressors appear more agreeable. *Journal of Personality*, *80*, 503-536.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In A. Elliot (Ed.), *Handbook of approach and avoidance motivation* (pp. 251-271). Mahaw, NJ: Lawrence.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*, 829-858.

Watanabe, E., & Ohtsubo, Y. (2012). Costly apology and self-punishment after an unintentional transgression. *Journal of Evolutionary Psychology*, *10*, 87-105.

Watts, D. P. (2006). Conflict resolution in chimpanzees and the valuable-relationships hypothesis. *International Journal of Primatology*, *27*, 1337-1364.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*, 2065-2072.

Zahavi, A., & Zahavi, A. (1997). *The handicap principle: A missing piece of Darwin's puzzle*.

Oxford: UK: Oxford University Press.

**Table 1.** Brain regions that were more active in the presence of costly apologies.

| Region | *Cluster size* | Peak MNI Coordinates x | y | z | *t* | Cluster *P* value (FWE-corrected) |
|---|---|---|---|---|---|---|
| Costly Apology > Non-costly Apology | | | | | | |
| ***Right TPJ***/Fusiform Gyrus | 6605 | 48 | −52 | 24 | 9.08 | < .001 |
| ***MPFC***/***Left TPJ***/Middle Temporal Gyrus | 17493 | −56 | −6 | −14 | 10.31 | < .001 |
| Right Middle Temporal Gyrus | 1548 | 58 | 0 | −20 | 9.39 | < .001 |
| ***Precuneus*** | 1796 | 10 | −50 | 36 | 8.32 | < .001 |
| Orbitofrontal Cortex (OFC) | 651 | 0 | 46 | −18 | 7.19 | < .001 |
| Right Dorsolateral Prefrontal Cortex (DLPFC) | 1044 | 36 | 8 | 36 | 5.73 | < .001 |
| Cerebellum | 225 | −4 | −54 | −44 | 6.08 | .038 |
| Costly Apology > No Apology | | | | | | |
| ***Right TPJ***/Fusiform Gyrus/Middle Temporal Gyrus | 8886 | 36 | −46 | −18 | 9.70 | < .001 |
| ***Left TPJ***/Fusiform Gyrus/Middle Temporal Gyrus | 8991 | −50 | −64 | 24 | 9.66 | < .001 |
| Left Middle Frontal Gyrus | 261 | −40 | 38 | −10 | 5.60 | .022 |

| | | | | | | |
|---|---|---|---|---|---|---|
| ***Precuneus*** | 1447 | 6 | −50 | 36 | 7.53 | < .001 |
| ***MPFC*** | 381 | −8 | 60 | 12 | 5.07 | .004 |
| Right DLPFC | 1532 | 16 | 42 | 44 | 7.04 | < .001 |
| Left DLPFC | 2360 | −26 | 26 | 54 | 7.51 | < .001 |
| OFC | 473 | −2 | 26 | −18 | 5.90 | .001 |
| Cerebellum | 635 | 8 | −70 | −26 | 5.61 | < .001 |
| Non-costly Apology > No Apology (No significantly different clusters were observed.) | | | | | | |

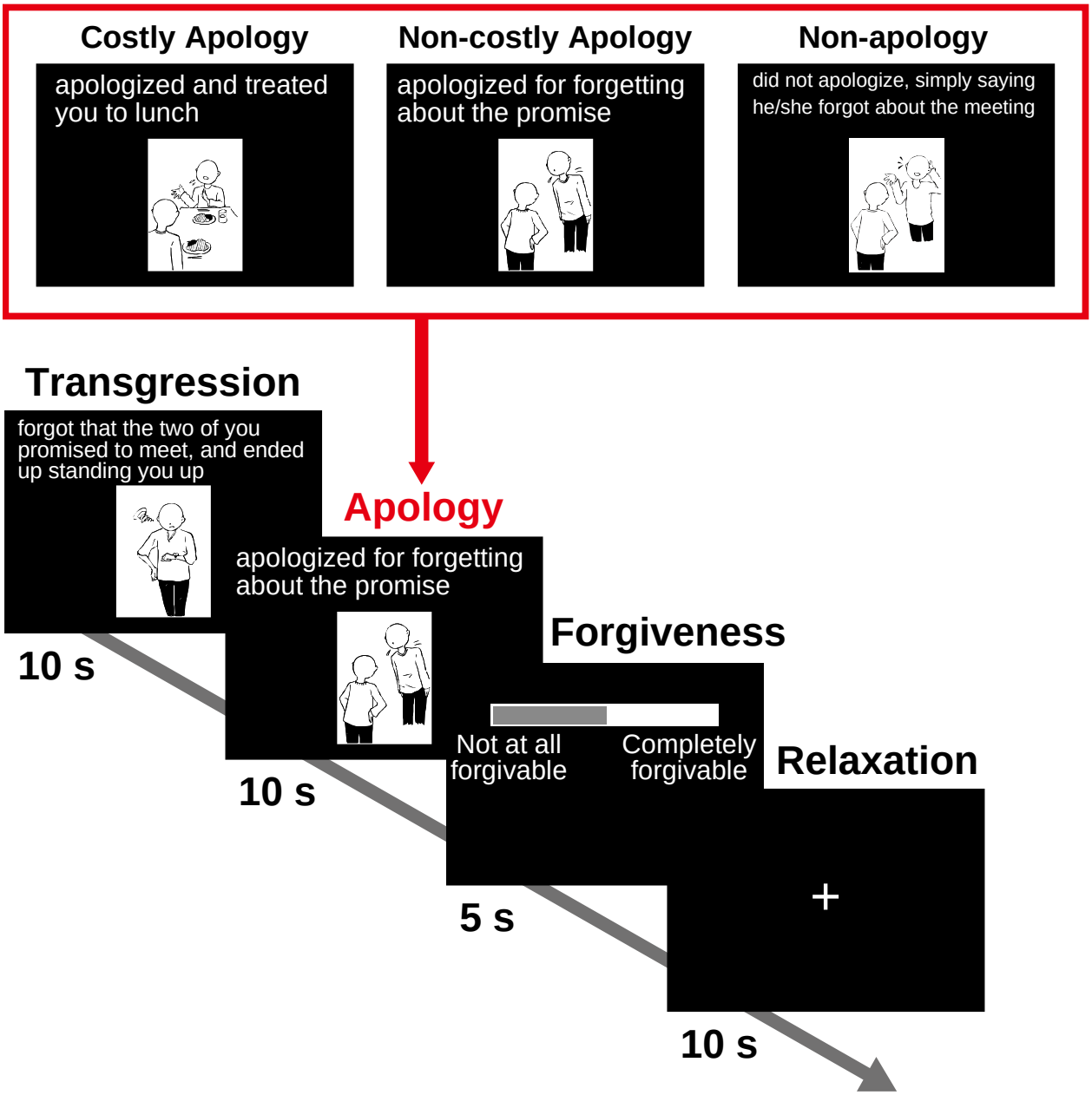


**Fig. 1.** Time Course of the Experiment. Each transgression scenario was followed by a reaction scenario, which was either a costly apology, a non-costly apology, or a non-apology. After observing the transgressor's reaction, participants determined how much they could forgive the transgressor. After a 10-second relaxation phase, the same procedure was repeated.

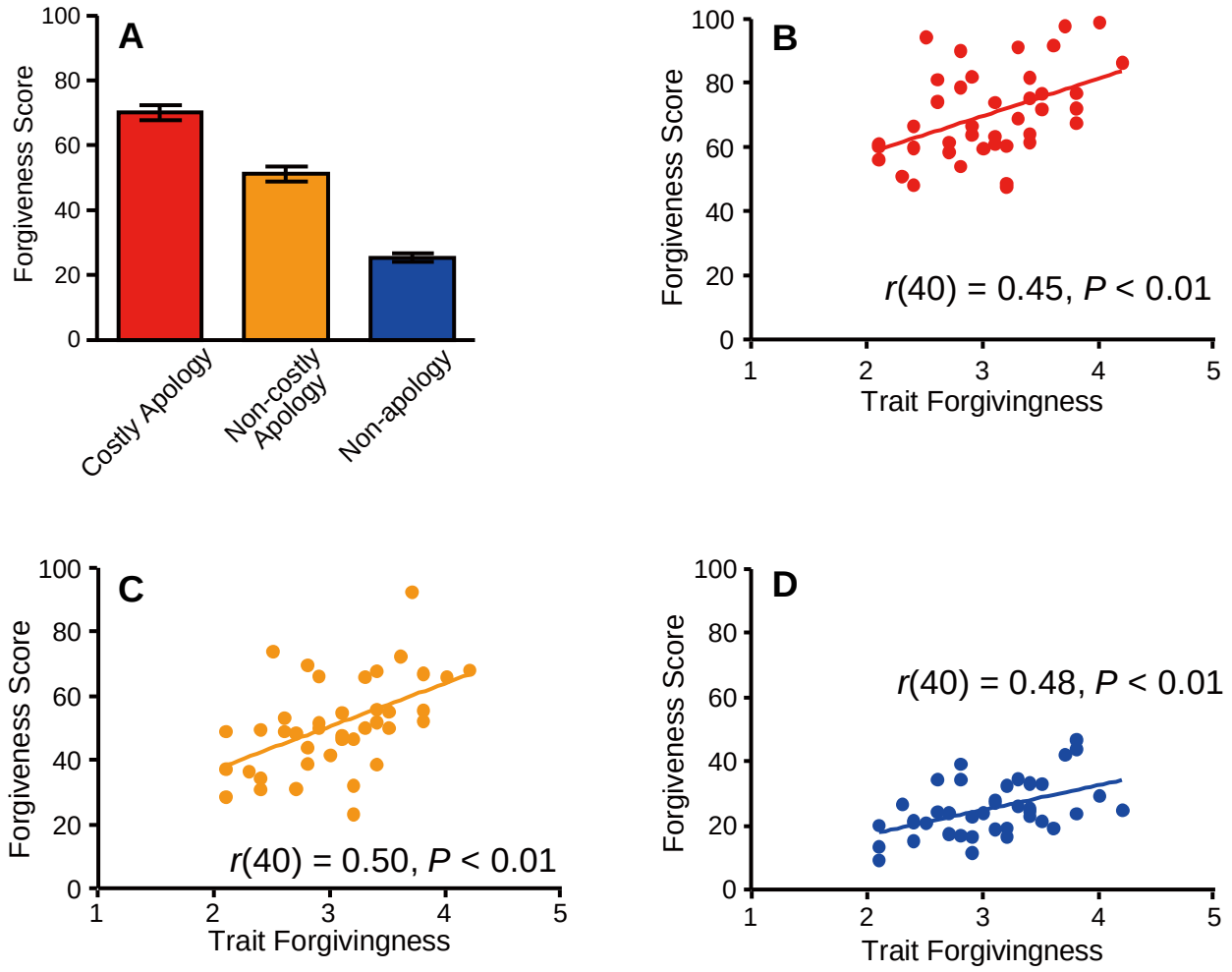


**Fig. 2.** Forgiveness and Trait Forgivingness. (A) Mean forgiveness as a function of reaction type (costly apology vs. non-costly apology vs. non-apology). The effect of reaction type was significant by a one-way ANOVA ($F_{2, 72} = 269.12$, $P < .001$, $\eta_p^2 = 0.88$), and post-hoc tests revealed significant differences in every pair of the three conditions ($t_{72} = 13.38$, 9.72, and 23.10 for costly apology vs. non-costly apology, non-costly apology vs. non-apology, and costly apology vs. non-apology comparisons, All *p*s < .001). (B) Scatter plot showing the positive correlation between trait forgivingness and forgiveness in the costly apology condition. (C) Scatter plot showing the positive correlation between trait forgivingness and forgiveness in the non-costly apology condition. (D) Scatter plot showing the positive correlation between trait forgivingness and forgiveness in the non-apology condition.

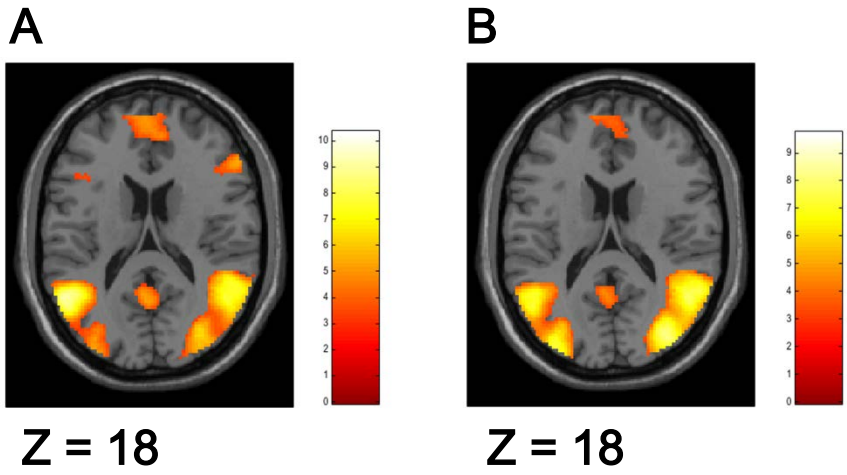


**Fig. 3.** Costly Apologies versus Both Non-costly Apologies and Non-apologies Produce Near Equivalent Brain Responses. (A) The responses of the bilateral TPJ, precuneus, and MPFC to costly apologies in comparison with non-costly apologies. (B) The responses of the bilateral TPJ, precuneus, and MPFC to costly apologies in comparison with non-apologies. Statistical significance thresholds were set at $p < .001$ (uncorrected) at the voxel level and $p < .05$ (familywise-error [FWE] corrected; whole brain) at the cluster level.

**Supplementary Online Materials**

# Costly Apologies Communicate Conciliatory Intention:

# An fMRI Study on Forgiveness in Response to Costly Apologies

Yohsuke Ohtsubo*, Masahiro Matsunaga, Hiroki Tanaka, Kohta Suzuki, Fumio Kobayashi, Eiji Shibata, Reiko Hori, Tomohiro Umemura, and Hideki Ohira

## Scenarios

We made the scenarios as brief as possible for the following reasons. First, we wanted to make the scanning session short to reduce the stress of participants. Second, we wanted to make participants' reading time as short as possible so that the BOLD responses would be less contaminated by activities related to individual differences in reading speed. Third, we expected that the number of participants who are discarded from the analyses due to excessive head movements could be minimized by shortening the scanning session. As shown in Table S1, the lengths of these scenarios were nearly equivalent: The transgression scenarios consisted of 17.7 letters (i.e. Japanese characters) on average, with the *SD* of 0.64. Scenario lengths ranged from 17 to 19 letters. The costly apology scenarios averaged 17.6 letters with an *SD* of 1.56 (ranging 15 to 20 letters). The non-costly apology scenarios averaged 15.0 letters with an *SD* of 2.05 (ranging 12 to 18 letters). The non-apology scenarios averaged 16.8 letters with an *SD* of 1.66 (ranging 14 to 19 letters). We were unsure if participants would find such brief costly apologies to be sincere than non-costly apologies. To confirm that the brief scenarios retain the effect observed in previous studies, we conducted an online pilot study.

**Table S1.** Scenarios in Japanese.

| | | |
|---|---|---|
| 1 | Transgression | あなたとの約束を忘れてすっぽかした |
| | Costly Apology | おわびにお昼をおごってくれて謝った |
| | Non-costly Apology | 約束を忘れたことを謝った |
| | No Apology | 謝らず、すっかり忘れていたといった |
| 2 | Transgression | あなたの好きな異性をみんなにばらした |
| | Costly Apology | うっかりを謝り、全員に口止めしてまわった |
| | Non-costly Apology | うっかりしていたといって謝った |
| | No Apology | 謝らず、気にすることじゃないといった |
| 3 | Transgression | あなたの物を勝手に別の人に又貸しした |
| | Costly Apology | すぐに相手から返してもらってきて謝った |
| | Non-costly Apology | どうしても断れなかったといって謝った |
| | No Apology | 謝らず、友達だからいいだろうといった |
| 4 | Transgression | お酒の席であなたの容姿をからかった |
| | Costly Apology | あなたに謝り、好きなお酒を断酒した |
| | Non-costly Apology | 悪気はなかったと謝った |
| | No Apology | 冗談だから気にするなといった |
| 5 | Transgression | あなたを入れずにLINEグループを作った |
| | Costly Apology | それに気づくと、すぐに慌てて謝りに来た |
| | Non-costly Apology | 声をかけ忘れたといって謝った |
| | No Apology | 忘れていただけでわざとじゃないといった |
| 6 | Transgression | 貸してあげた物を乱暴に扱った壊した |
| | Costly Apology | 壊した物を買いなおして謝りに来た |
| | Non-costly Apology | 壊すつもりはなかったといって謝った |
| | No Apology | 調子が悪いとだけいって返した |
| 7 | Transgression | みんなで遊びにいくのに、誘わなかった |
| | Costly Apology | おわびに別の企画をたててくれた |
| | Non-costly Apology | 忙しいと思っていたといって謝った |
| | No Apology | 謝らず、興味ないと思ったといった |
| 8 | Transgression | あなたのスマホで下品な書き込みをした |
| | Costly Apology | なんでもいうことをきくといって謝った |
| | Non-costly Apology | 冗談のつもりだったといって謝った |
| | No Apology | 謝らず、冗談だと誰でもわかるといった |
| 9 | Transgression | あなたの秘密の悩みをみんなにばらした |
| | Costly Apology | 徹夜で悩みの解決法を調べて、謝りにきた |
| | Non-costly Apology | 気にしないと思ったといって謝った |
| | No Apology | 謝らず、たいした悩みじゃないといった |
| 10 | Transgression | あなたとの待ち合わせに1時間も遅刻した |
| | Costly Apology | 遅刻したことを謝り、おやつをおごってくれた |
| | Non-costly Apology | 寝坊してしまったといって謝った |
| | No Apology | 寝坊したといって特に謝らなかった |

## Pilot Study

### Procedure

Participants were recruited through an online survey service provided by Cross Marketing Inc., Japan. Eligibility for participation in this study was limited to individuals currently enrolled in either a university, college or vocational school. Three hundred participants (150 females and 150 males, 20.4 ± 1.17 years-old) completed the online survey. The survey consisted of the following four sections. The first section asked participants' demographic information, including sex and age. The second section consisted of the Japanese Trait Forgivingness Scale (J-TFS) (Ohtsubo, Yamaura, & Yagi, 2016). The J-TFS is a 10-item single-factor measure of individual differences in the inclination to forgive others. Sample items in the original English version (Berry, Worthington, O'Connor, Parrott, & Wade, 2005) include "I can forgive a friend for almost anything," and "I feel bitter about many of my relationships" (reverse-coded item). These items were accompanied by a 5-point scale (1 = "completely disagree" to 5 = "completely agree"). The 10-item scores were aggregated to obtain the trait forgivingness score (Cronbach's $\alpha = .76$). The third section consisted of the Japanese version of the Revised UCLA Loneliness Scale (Moroi, 1991). This measure is irrelevant to the present purpose. The fourth section consisted of three randomly-chosen transgression scenarios. Each scenario was accompanied by one of the three perpetrator reactions (costly apology, non-costly apology, or non-apology). The reaction scenarios were chosen in a semi-random manner whereby the three types of reactions were always included in each participant's survey. The order of the three reactions was randomized. The fourth section was the main part of the pilot study. The fourth section comprised an experiment involving a single within-participant factor with three levels (reaction type: costly apology vs. non-costly apology vs. non-apology).

Each transgression scenario was followed by three items designed to assay participants' initial negative feeling towards the transgressor: (1-1) *How much would you be angry at your friend if this really happened?* (1-2) *How much would you be irritated at your friend if this really happened?* (1-3) *How likely do you think you would be to dissolve the relationship with your friend if this really happened?* These three items were accompanied by a 5-point scale (0 = "not at all" to 4 = "very much"). After responding to these items, participants were exposed to the friend's reaction, and subsequently responded to the following seven items. All seven items started "If your friend in fact reacted this way…" (2-1: post-reaction anger) *How much would you be still angry at your friend?* (2-2: forgiveness) *How much could you forgive your friend?* (2-3: perceived exploitation risk) *How likely do you think your friend is to commit the same transgression again?* (2-4: intentionality of the transgression) *How probable do you think it that your friend intentionally committed the transgression?* (2-5: transgressor valuation of the relationship) *How much do you think your friend values the relationship with you?* (2-6: empathy) *How empathetic would you feel towards your friend?* (2-7: sincerity) *How sincere would you think your friend's reaction is?* These seven items were accompanied by a 5-point scale (0 = 'not at all' to 4 = 'very much').

**Results and Discussion**

The purpose of the pilot study was to confirm the validity of the scenarios for use in the main study. In particular, we examined whether brief descriptions of costly apologies are sufficient to induce perceived sincerity, and thereby, intention-reading. As is discussed below, this online study confirmed the validity of the scenarios.

In the pilot study, participants were first asked about their feelings towards a transgression without being informed of the transgressor's reaction. The three pre-reaction feeling items (1-1, 1-2, and 1-3) were internally consistent (Cronbach's $\alpha$ = .88, .86, and .86 for

the costly apology, non-costly apology, and non-apology scenarios, respectively), and were thus aggregated into a variable we call pre-reaction feeling. As expected, pre-reaction feeling did not significantly vary across the three conditions, although there was a marginally significant omnibus effect: $M_{\text{Costly Apology}} \pm SD = 2.85 \pm 1.10$, $M_{\text{Non-costly Apology}} \pm SD = 2.97 \pm 1.06$, and $M_{\text{Non-apology}} \pm SD = 2.84 \pm 1.09$, $F_{2, 598} = 2.39$, $p = .092$. Therefore, before knowing the transgressor's reaction, there were no systematic differences in participants' feelings towards the event.

The descriptive and test statistics associated with the seven post-reaction items are summarized in Table S2. Two items most relevant to intention-reading are perceived sincerity (2-7) and perceived exploitation risk (2-3). As reported in the main text, perceived sincerity was significantly greater in the costly apology condition than in the non-costly apology condition ($t_{598} = 2.71$, $p = .007$) and the non-apology condition ($t_{598} = 10.16$, $p < .001$). Perceived exploitation risk was significantly lower in the costly apology condition than in the non-costly apology condition ($t_{598} = 2.96$, $P = .003$) and the non-apology condition ($t_{598} = 3.48$, $p < .001$). These results suggest that costly apologies more effectively communicated the transgressor's sincere intention and reduced the victim's suspicion about the transgressor's exploitative intention. It is conceivable that valuation of the transgressor (2-5) is also relevant to intention-reading; if the victim believes the transgressor values the relationship with him/her, it is reasonable for the victim to expect that the transgressor will not exploit him/her again. Accordingly, participants considered that the transgressor values the relationship more when they received costly apologies than non-costly apologies ($t_{598} = 2.77$, $p = .006$) and non-apologies ($t_{598} = 8.26$, $p < .001$). In addition, costly apologies more effectively reduced participants' post-reaction anger than non-costly apologies ($t_{598} = 3.69$, $p < .001$) and non-apologies ($t_{598} = 10.29$, $p < .001$).

A slightly unexpected pattern was observed for forgiveness (2-2). Although the direction of the mean differences was in line with our prediction (i.e. costly apology > non-costly apology

> non-apology), the difference between the costly apology and non-costly apology conditions failed to reach the level of conventional significance ($p$ = .05). However, as we reported in the main text (see Fig.2A), in the fMRI study, forgiveness was significantly greater in the costly apology condition than in the non-costly apology condition. These contradictory results might have been due to lower reliability of this online study, in which participants only observed three scenarios, while participants in the fMRI study observed all 30 scenarios. In fact, the correlation coefficients between forgiveness scores and trait forgivingness score were smaller in the online study ($r$s = .16, .27, and .14 for the costly apology, non-costly apology and non-apology conditions, respectively, all $p < .05$) than in the fMRI study (comparable $r$s = .45, .50, and 53; see Figs.2B to 2D). Another unexpected pattern was observed for empathy towards the perpetrator (2-6). Similar to the results of forgiveness, the results indicated a non-significant difference between the costly and non-costly apology conditions, whereas the differences between these two conditions and the non-apology condition were significant.

**Table S2.** Descriptive and Test Statistics of the Online Pilot Study.

| | Costly Apology | Non-costly Apology | Non-apology | $F_{2, 598}$ | $p$ | $\eta_p^2$ |
|---|---|---|---|---|---|---|
| Post-anger | $2.32 \pm 1.12^a$ | $2.62 \pm 1.17^b$ | $3.15 \pm 1.16^c$ | 54.30 | < .001 | 0.15 |
| Forgiveness | $3.64 \pm 1.11^a$ | $3.54 \pm 1.08^a$ | $3.03 \pm 1.15^b$ | 35.05 | < .001 | 0.10 |
| Exploitation Risk | $2.86 \pm 1.04^a$ | $3.07 \pm 1.08^b$ | $3.32 \pm 1.08^c$ | 20.74 | < .001 | 0.06 |
| Intentionality | $2.19 \pm 1.10^a$ | $2.22 \pm 1.13^a$ | $2.48 \pm 1.15^b$ | 11.96 | < .001 | 0.04 |
| Valuation | $3.39 \pm 1.02^a$ | $3.21 \pm 1.01^b$ | $2.84 \pm 1.01^c$ | 35.35 | < .001 | 0.11 |
| Empathy | $3.10 \pm 1.09^a$ | $3.12 \pm 1.10^a$ | $2.71 \pm 1.05^b$ | 19.36 | < .001 | 0.06 |
| Sincerity | $3.44 \pm 1.04^a$ | $3.25 \pm 1.05^b$ | $2.51 \pm 0.98^c$ | 92.12 | < .001 | 0.24 |

*Note.* Different superscripts (a, b, and c) indicate significant difference between the means associated with different superscripts.

Another interesting result was associated with intentionality (2-4). This item asked participants to rate how much they supposed the transgressor intentionally committed the transgression. Of note, this form of mind-reading is very different from imagining the *prospective* intention of the transgressor (i.e., whether the perpetrator will exploit the victim again). It is retrospective; it concerns how much the victim considers that he/she was intentionally harmed in the past. Interestingly, although costly apologies had significant effects on prospective intention-related measures (i.e. sincerity and exploitation risk), there was no significant difference in (retrospective) intentionality between the costly apology and non-costly apology conditions ($t_{598}$ = 0.46, $p$= .646). This may be due to the fact that people appear to make rapid inferences about the intentionality of others' past behaviors based on information such as whether the transgressor had the *ability* to exploit the victim and/or held the *belief* that the behavior would lead to the victimization (Malle & Knobe, 1997).

In sum, the results of pilot study confirmed the validity of the materials that were used in the fMRI study. Most importantly, we have shown that the brief descriptions of costly apologies are sufficient to convey sincere intention.

## References


Berry, J. W., Worthington, E. L. Jr., O'Connor, L. E., Parrott, L., III, & Wade, N. G. (2005). Forgivingness, vengeful rumination, and affective traits. *Journal of Personality*, *73*, 183-225.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101-121.

Moroi, K. (1991). *Kaitei* UCLA *kodokukan syakudo no jigensei no kentou* [Examination of Revised UCLA Loneliness Scale's dimensions]. *Shizuoka University Repository*, *42*, 23-51.

Ohtsubo, Y. Yamaura, K., & Yagi, A. (2015). Development of Japanese measures of reconciliatory tendencies: The Japanese Trait Forgivingness Scale and the Japanese Proclivity to Apologize Measure. *Japanese Journal of Social Psychology*, *31*, 135-42.