DR. JEROEN VAES (Orcid ID : 0000-0003-2256-2453)

Article type : Research Article

Corresponding Author Email ID: jeroen.vaes@unitn.it

RUNNING HEAD: Tethered humanity

Tethered humanity: Humanizing self and others in response to interpersonal harm

Total word count: 10.762 words

Abstract

We propose an integrative framework in which the self-image of perpetrators and victims in interpersonal conflictual relationships becomes tethered to each other. To the extent that both parties stop seeing themselves as fully human when interpersonal harm is inflicted, we theorize that this state motivates attempts at restoration. Specifically, we propose the *tethered humanity hypothesis* which ascertains that perpetrators or victims can reclaim their full human status by re-establishing the humanness of the opposing party. This hypothesis was analyzed from both the perspective of the perpetrator (Study 1-3) and the victim (Study 4). In all studies, results indicated that perpetrators or victims self-dehumanize in response to interpersonal harm and managed to regain a full human status when re-humanizing the opposing party. Interestingly, this process was interrupted when the opposing party did not accept attempts at reconciliation. Our findings offer a new interpersonal perspective on the implications of immoral behavior.

Keywords: Self-dehumanization, tethered humanity hypothesis, interpersonal conflict, immoral behavior

People frequently engage in ego-protective responses to justify their own wrongdoing, seeking to distance themselves from their own immoral behavior (Bandura, 1999). Those, instead, who suffer immoral treatment might seek revenge to restore their damaged self-image (Jackson, Choi, & Gelfand, 2019). From this perspective, interpersonal transgressions should routinely lead to an ongoing breakdown of relations (e.g., Goldstein, Davis, & Herman, 1975; Martens, Kosloff, Greenberg, Landau, & Schmader, 2007). Yet, research has begun to demonstrate that people sometimes take responsibility for their moral transgressions and that perpetrators and victims are motivated to behave in ways that restore their own moral self-concept (e.g., Bastian, Jetten, Chen, Radke, Harding, & Fasoli, 2013; Jordon, Mullen, & Murnighan, 2011; McCullough, Root, Tabak, & Witvliet, 2009; Sachdeva, Iliev, & Medin, 2009; Tetlock, Kristel, Elson, Green, & Lerner, 2000; Zhong & Liljenquist, 2006). In this paper, we examine a novel perspective on how people can re-establish their moral self-image in conflictual interpersonal relations. We suggest that through a recognition of their immoral behavior, perpetrators become tethered to their victims and in their attempt to revalidate and empower themselves, victims can

become bound to their perpetrators. To the extent that they acknowledge the harm they inflicted, perpetrators may view their self as less human, yet it may only be through re-establishing the full humanness of their victims, they can restore their lost humanity. Victims of interpersonal harm also self-dehumanize because they feel degraded and might – through processes of forgiveness – rehumanize the perpetrator and themselves. This perspective offers new insight into the power relations that exist between perpetrators and their victims: just as victims may need perpetrators to acknowledge their wrongdoing, perpetrators are sometimes also at the mercy of their victim's willingness to reconcile.

Self- and other dehumanization

There is no kind of interpersonal relationship in which people do not get harmed. Whether it might be a lie, an insult or the social exclusion of the other, these interpersonal offenses have psychological and relational consequences for the transgressor and the victim (e.g., Baumeister, Stillwell, & Wotman, 1990) and lead to changes in perception of both self and other (Bastian, Jetten, & Haslam, 2014; Bastian et al., 2013). Perpetrators, when treating others harmfully may respond to their own behavior in a variety of ways. One group of reactions are self-serving or ego-protective responses that aim to avoid the possibility of self-sanctions when behaving immorally (e.g., Bastian & Loughnan, 2017). Such strategies of moral disengagement restructure one's conduct into benign or worthy behavior through the use of euphemisms, the diffusion or displacement of responsibility and the attribution of blame to the victim, to name but a few (Bandura, 1999).

In contrast to this ego-protective response, there is a growing body of work showing that people often take responsibility for their moral transgressions (e.g., Jordon, Mullen, & Murnighan, 2011; Sachdeva, Iliev, & Medin, 2009). This possibility has also been explored within the dehumanization framework, finding that people self-dehumanize in response to their own moral transgressions (Bastian, Jetten & Radke, 2012; Bastian et al., 2013). Self-dehumanization results then from the recognition one's actions are immoral and have caused harm to others which cannot be completely justified otherwise. Given the close link between immoral conduct and dehumanization (Bastian, Laham, Wilson, Haslam, & Koval, 2011; Brandt & Reyna, 2011; Gray, Young, & Waytz, 2012), acting in immoral ways diminishes the extent to

which people feel they possess human qualities relative to when they acted with moral regard.

The self-perceptions of victims of interpersonal harm are affected too. When treated with disrespect, contempt or being socially excluded by others, victims might feel degraded and ultimately less human. In the context of social ostracism, Bastian and Haslam (2010) demonstrated that compared to inclusive or everyday interactions, being socially excluded led victims to see themselves as less human. In addition, other forms of social indignities that undermine basic elements of personhood, like identity and status have been shown to lead to self-dehumanization among victims (Bastian & Haslam, 2011).

While victims and perpetrators of interpersonal harm adjust their self-perception in human terms, they often dehumanize the opposite party as well. Perpetrators might dehumanize their victims to avoid self-sanctions (Bandura, 1999). Once dehumanized, they are no longer seen as persons with feelings, intentions, and thoughts justifying the fact that one mistreated them. Still, unlike the self-dehumanization of the perpetrator, no harm or mistreatment is necessary to dehumanize the victim. Especially in conflictual relationships, the victim may be dehumanized independently of the harm caused by the perpetrator. This prediction is consistent with research showing that priming generic violence (Delgado, Rodriguez-Pérez, Vaes, Leyens, & Betancor, 2009) or inducing negative feelings such as anger, contempt, or disgust (Bastian, Denson, & Haslam, 2013; Buckels & Trapnell, 2013; Harris & Fiske, 2006) are sufficient to elicit or intensify the dehumanization of others.

Even though studied to a lesser extent, victims of interpersonal harm might dehumanize their perpetrators. In this regard, Bastian and Haslam (2010) asked participants to either recall an event in which they were socially excluded or to play Cyberball (Williams et al., 2000), a ball tossing game in which participants are excluded by other players. In both cases, results indicated that the perpetrators of social exclusion were seen as having fewer human qualities by those who they excluded. Overall, these studies suggest that perpetrators dehumanize their victims as it might help them to justify their immoral acts, but when they acknowledge that harm, they will also self-dehumanize. Victims of interpersonal harm show a very similar reaction as they tend to dehumanize their perpetrators and self-dehumanize in response to the treatment they received.

Tethered humanity

Most of the research on interpersonal conflict management has either focused on the perspective of the victim studying, for example, processes of forgiveness (e.g., Fehr, Gelfand, & Nag, 2010; Wade, Hoyt, Kidwell, & Worthington, 2014) or analyzing reconciliating actions on the side of the perpetrator, like apologies (e.g., Ohbuchi, Kameda, & Agarie, 1989; Schumann, 2018). In the current article, instead, we propose an integrative framework in which perpetrators and victims in interpersonal conflictual relationships become tethered to each other. To the extent that both parties lose their moral self-image when interpersonal harm is inflicted, they will be motivated to restore their self-image. In the former paragraphs, we described how people tend to self-dehumanize when they either perpetrate harm and acknowledge that harm or when they become the victims of such mistreatments. In both cases, we argue that this response represents a motivational state in which people seek to reclaim their full human status. Feeling human satisfies fundamental needs, like identity (Demoulin et al., 2009; Paladino et al., 2004) or existential needs (Goldenberg et al., 2001; Goldenberg, Heflick, Vaes, Motyl, & Greenberg, 2009; Vaes, Heflick, & Goldenberg, 2010) and is driven by the perception that human traits are deeply rooted and central to one's personal identity (Haslam, Bain, Douge, Lee, & Bastian, 2005). Therefore, perpetrators and victims alike, when they have tarnished their own self-image in this way, will be motivated to reclaim it.

This brings us to the key prediction in this paper – the *tethered humanity hypothesis*. One way for perpetrators or victims to reclaim their full human status, is to re-establish or at least increase the perceived humanity of the opposing party. Both parties in the interpersonal conflict have lost a valued state motivating them to change their views of the opposing party. From the perspective of the perpetrators, we argue that when perpetrators acknowledge their wrongdoing and self-dehumanize as a result, their self-concept becomes tethered to that of their victims. Therefore, reclaiming self-humanity might be impeded if the other party refuses to reconcile. To reclaim their own humanity, they must restore that of their victims as well, meaning that when one's victim refuses to reconcile (e.g., by not accepting their apology or refusing to forgive them) it undermines the perpetrator's capacity to restore their victim's humanity, and therefore their own humanity. In a similar vein, victims may be able to restore their tarnished self-image if they can rehumanize their perpetrator,

perhaps by forgiving them. If the perpetrator refuses to apologize or to express sincere remorse, this impedes the victim's ability to rehumanize the perpetrator and therefore the self.

We make two early observations about our predictions. First, the predicted pattern of results could be driven by basic principles of cognitive dissonance theory (e.g., Festinger, 1957; Harmon-Jones & Mills, 1999; Kenworthy, Miller, Collins, Read, & Earleywine, 2011). For instance, an inconsistency between my desire to be human and my currently dehumanized state. However, dissonance theory cannot account for why it is that self- and other-humanity becomes tethered. Furthermore, we see the motivation to reclaim one's humanity as driven by several significant needs and drives (such as those reviewed above), which extend beyond a straightforward inconsistency in self-perception.

Second, we acknowledge that alternative routes to restoring one's moral and therefore human self-image exist. Research has shown that acting in accordance with one's moral principles in private (Jordon, Mullen, & Murnighan, 2011) or subconsciously engaging in acts of self-cleansing (Zhong & Liljenquist, 2006) might have a similar effect. Also, people appear to be able to reclaim humanity when their prosocial behavior is directed towards others, rather than the victims of their actions (e.g., Bastian et al, 2013). Our research also contrasts with recent suggestions that self-dehumanization resulting from engaging in unethical behavior may facilitate future unethical behavior (Kouchaki, Dobson, Waytz, & Kteily, 2018). While Kauchaki and colleagues focused on self-regulation failures in the moral domain (i.e., succumbing to the temptation to lie or cheat for money) without any negative consequences for known others, in the current work we focus on immoral actions that cause harm to an identifiable victim. It is in interpersonal conflictual relationships, where both the perpetrator and the victim are known, that we expect self-dehumanization to lead to compensatory reactions to restore both the humanness of the self and the opposing party.

Overview of studies

Across four studies we test the tethered humanity hypothesis. Specifically, we predict that both perpetrators and victims self-dehumanize in response to interpersonal harm, and this change in self-perception motivates attempts to regain a full human status. One way to obtain such status is through the re-humanization of the

other party. The first three studies analyzed this hypothesis from the perspective of the perpetrator, while the last study measured participants reactions in the victim role. In order to test this hypothesis, we needed to create a conflictual interpersonal context in which harm was inflicted. In the first two studies, participants were asked in a recall paradigm to think about a good friend (Study 1) or an "enemy" (Study 1 and 2) and imagine themselves in a number of situations in which they could behave aggressively towards that person. In the latter two studies, conflict and interpersonal harm were created in the laboratory asking participants to engage in rude behavior during an online chat (Study 3) or exposing them to the experience of social ostracism (Study 4). Each study tested specific aspects of the tethered humanity process. In Study 1, comparing participants' reaction to a friend or an "enemy", we expected participants to inflict harm on an "enemy", but not on a friend, resulting in a process of self-dehumanization only in the "enemy" condition. After an interval, and giving participants the possibility to help the target, we expected participants to rehumanize the self and the other in the enemy condition restoring a full human status. Study 2 was specifically designed to test some of the conditions that trigger both the perpetrator's tendency to self-dehumanize and the tethered humanity effect. Specifically, this study aimed to demonstrate that self-dehumanization engages attempts at re-humanization of both the self and the other. In the last two studies, we aimed to show that rehumanizing the self is most likely when the other party accepts the attempts for reconciliation. In Study 3, participants were led to engage in the role of a rude or friendly conversation partner. In the rude condition, all perpetrators were instructed to apologize for their conduct, but their apology was either accepted or not by the victim. In Study 4, instead, participants were included or socially excluded in a Cyberball game. When excluded they interacted with one of the perpetrators who either sincerely apologized or not for his or her conduct during the game. In both studies, the dehumanization of the self and the other were measured immediately after the transgression and after the attempt for reconciliation. We expected participants, victims and perpetrators alike, to self-dehumanize when harm was inflicted and to rehumanize themselves and the other party only when the attempt for reconciliation was accepted. Indeed, when this effort for reconciliation is interrupted, victims or perpetrators may be unable to rehumanize the other and therefore themselves.

Across the four studies we introduced two different humanness measures that incorporate concepts that are related to the two dimensions of dehumanization. Mechanistic dehumanization involves the perception of others as more machine-like, denying human nature attributes, while the other occurs when we have an animalized view of others, denying them uniquely human traits, and is referred to as animalistic dehumanization (Haslam, 2006). In the current set of studies, however, our primary focus is on the perception of the self or the other as lacking or having lost qualities that are considered human, including aspects of both senses of humanness. Even though previous work has shown that these two dimensions might operate independently, research on self- and other dehumanization has shown that in the case of interpersonal maltreatment both forms of dehumanization tend to co-occur (Bastian et al., 2012; 2013). Self-evaluation concerns and sometimes mood were also measured, to reassure it is the loss of humanity and not simply the loss of general positivity that drives people's need to restore their full humanity and re-humanize the other. All studies adhered to the ethical guidelines specified in the APA Code of Conduct as well as the national ethics guidelines in which the studies were conducted. The data of all studies can be found on the Open Science Framework (https://osf.io/v6bj9/)

Study 1

In Study 1, participants were asked to recall and describe their relationship with a good friend or an "enemy" (i.e., somebody with whom they had a difficult or conflicting relationship in the past six months). They were then asked to imagine a number of situations in which they could behave aggressively towards the target person and describe the self and the other in human terms when reflecting back on how they behaved. Given that we expected participants to behave aggressively only in the "enemy" condition, we expected to observe self-dehumanization in the enemy but not in the friend condition (Hp1). In addition, and given the conflictual nature of their relationship, we expected participants to dehumanize the enemy, but not their friend (Hp2). Finally, in a second phase participants were asked to imagine a number of situations in which they could help the target and again describe the self and the other in human terms when reflecting back on how they behaved. In this phase, we expected to observe the tethered humanity effect showing that in the enemy condition

both the self and the other were seen as more human compared to the first measurement (Hp3).

Method

Participants

A power analysis using PANGEA (Power Analysis for GEneral Anova designs, Westfall, 2016) showed that a sample of 100 participants in each condition would be sufficient to detect the expected three-way interaction with a power of .90, even if this effect would be small (d = .30). On the basis of this analysis, we decided to gather a total sample of 250 participants (130 male, 119 female and 1 non-defined) on MTurk who were paid 1$ for their participation. Participants age ranged between 20 and 72 years old (M_{age}=37.57, SD=11.75). The majority of the sample identified as Caucasian (77.6%), while minorities were represented as follows: African American (8.8%), Asian (7.2%), Native American (3.2%), and other ethnicity (3.2%). A total of 15 participants were discarded from further analyses; eleven because they did not respond to all questions and four because they did not describe a credible friend or "enemy".

Procedure

Participants responded to a questionnaire in which they were first asked to think about a friend or an "enemy" depending on condition. A friend was defined as "a person with whom you have or have had a friendly relationship of trust and mutual respect", while an "enemy" was defined as "a person with whom you have or have had a difficult and conflictual relationship". In both cases, we asked participants to briefly describe the relationship they have (had) with the person and emphasized that the relationship had to be ongoing or could not date back to more than 6 months. After their brief description they were asked to report the gender and age of the target person.

In order to give participants the possibility to inflict personal harm on the target person they were asked to imagine themselves in three hypothetical situations (i.e., engaging in the social exclusion of the other, talking badly about the person behind his/her back, and a physically violent reaction in a sports game) in which they could behave aggressively towards the target person. For each situation they were asked to indicate how likely they would engage in the described behavior on a 7-point scale (1=*not at all* to 7=*very much*). After these scenarios, they were asked to imagine

finding themselves together with the other person in a psychological experiment in which they could determine for how long the target person had to keep his/her hands in cold water, a procedure known to induce pain (Bastian, Jetten, & Ferris, 2014). Participants were requested to indicate both the duration (0=*0 sec.* to 5=*50sec.*) and the temperature (1 = *10°C/50°F* to 6 = *0°C/32°F*) of the water. These indices were multiplied to obtain an index of the pain that was inflicted on the target person.

Participants then rated themselves first and then the target person on eight items adapted from Bastian and Haslam (2010) assessing the attribution of Human Nature (four items; for example, "I felt like I was open minded, like I could think clearly about things") and Human Uniqueness (four items; for example, "I felt like I was refined and cultured"). While responding on a 7-point scale (1=*not at all* to 7=*very much*), participants were explicitly requested to think back to how they behaved in the four hypothetical situations. These instructions were given to make it hard for participants who previously indicated they would behave aggressively toward the target person, to disengage from the harm they inflicted.

Some additional measures were taken including a general self-esteem scale (Rosenberg, 1965). This scale is made-up of 10 items and expressed what participants thought about themselves in that specific moment on a 6-point scale (1=*not at all* to 6=*very much*). Immediately afterwards participants were presented with four new hypothetical situations in which they had to imagine behaving prosocially towards the target person. In the first three situations (e.g., render money that the person lost, helping the person when hurt, feeling bad for the person's distress) participants indicated how likely they would help on a 7-point scale (1=*not at all* to 7=*very much*). In the fourth situation participants were asked to imagine finding themselves in a psychological experiment in which they could divide $10 between themselves and the other person. Participants were informed that their choice had to be accepted by the other person and that they could make their decision anonymously and privately. Under these circumstances, how much money would you be willing to give to the other person between $0 and $10? Finally, participants were asked to judge the self and the other again on the same humanness scale as the one that was used the first time, now imagining how they behaved in the helping situations.

In the end, participants were thanked for their participation and fully debriefed.

Results and discussion

We will only report the main analyses here and refer to the Supplemental Online Material (SOM) for a description of the analyses of all the variables. First of all, and as we expected, participants indicated to be willing to aggress, but refrained from helping their enemy compared to their friend. As such, we managed to create a conflictual (enemy) vs. a non-conflictual (friend) condition in which participants indicated a willingness to harm the other only in the former, but not in the latter condition. We analyzed participants self and other humanness judgements in two steps.

To verify participants initial tendency to dehumanize both the self and the other, we examined differences in self and other humanness judgments as a function of condition directly after the hypothetical aggressive situations only. Pairwise comparisons supported Hp1 showing that participants dehumanized the self more in the enemy compared to the friend condition, $F(1,232)=21.73$, $p<.001$, $\eta^2_p=.09$. In line with Hp2, pairwise comparisons indicated that participants attributed less humanness to the enemy compared to themselves, $F(1,232)=224.56$, $p<.001$, $\eta^2_p=.49$, while no differences in humanness between the self and the other were observed in the friend condition, $F(1,232)=.31$, $p=.58$, $\eta^2_p=.001$ (see Figure 1). Importantly, all these effects remained significant when controlling for participants' self-esteem suggesting that the effect was independent of the general valence in participants' self-regard (all p's$<.001$).

The second type of analyses compared the self and other humanness judgments over time in the two conditions separately to verify the existence of a re-humanization effect. In the friend condition, both the self and the friend were attributed high levels of humanness at both moments of measurement, but their humanness even further increased after the hypothetical helping situations, $F(1,128)=13.41$, $p<.001$, $\eta^2_p=.10$. While this increase in humanness was unexpected, it is likely that imagining truly helping one's friend makes people attribute even more humanness to both the self and that friend at Time 2 compared to imagining not aggressing against him or her at Time 1.

In the enemy condition, and confirming the tethered humanity hypothesis (Hp3), both the self and the enemy were attributed more humanness at Time 2 compared to the first measurement ($F(1,104)=121.88$, $p<.001$, $\eta^2_p=.54$ and

$F(1,104)=89.49$, $p<.001$, $\eta_p^2=.46$, for self and other judgments respectively). Overall, these results indicate that when people imagine inflicting harm in a conflictual interpersonal relationship they dehumanize both the self and their victim. In an attempt to restore this loss in humanity, the perpetrators humanize their victims and themselves.

Study 2

Study 1 provided support for our predictions drawn from the tethered humanity hypothesis. Participants dehumanized the self only in the enemy condition when they reflected back on their immoral conduct towards a conflictual other. This loss in humanity was successively restored rehumanizing the victim. Study 2 was specifically designed to unravel some of the conditions that allow this effect to occur. Four "enemy" conditions were created in which participants either imagined aggressing against the target or not and in which, in a second moment, they were asked to imagine helping the target or not. Replicating Hp2 of the first study, we expected participants to dehumanize the disliked other in all conditions given that hostile feelings towards the target were ever-present.

Previous work on self-dehumanization has emphasized the importance of recognizing one's conduct as immoral in triggering a process of self-dehumanization (Bastian et al., 2012; 2013). Therefore, we expect that self-dehumanization will only occur when people not only imagine a disliked other, but are also encouraged to treat that person immorally (Hp1). In an attempt to better understand this process, we tested the role of moral engagement. Immoral conduct will most likely lead to self-dehumanization to the extent that people morally engage with their own behavior, acknowledging that they behaved immorally and taking responsibility for their own conduct. In order to test this hypothesis, we assessed whether participants' tendency to morally engage with their own immoral conduct mediated their tendency to self-dehumanize (Hp4).

Creating a condition in which self-dehumanization occurs and one in which it does not, allowed us to verify whether this process enables the occurrence of the tethered humanity effect. We argue that as self-dehumanization represents the frustration of a basic motivation to see oneself as human, which in turn is expected to trigger a desire to change this state of affairs. When people do not self-dehumanize, they may continue to view their victim as dehumanized whilst maintaining a

perception that their own humanity has been untarnished. Therefore, perpetrators are only expected to rehumanize both the self and the other when they self-dehumanized in the first place (Hp3a). In addition, we also argue that even when helping one's victim is not possible, the motivation to regain humanity triggered by the perpetrator's initial tendency to self-dehumanize, should lead to a shift in the perception of one's victim, re-humanizing them such that one can also re-humanize the self. Therefore, we manipulated the presence of the hypothetical helping situations. If self-dehumanization triggers a general motivation to regain humanity, then we should see the re-humanization of the other even when other avenues are not available; such as helping one's victim (Hp3b).

The main dependent variable was changed in Study 2 introducing a trait measure of humanness. In addition, Study 2 controlled for the gravity of conflicts that were self-generated by participants both through external judges and participants self-judgments. Also, along with self-esteem, a mood measure was added and apart from the humanness traits participants described both the self and the other on traits that had a clear positive or negative valence, but were not well-defined on the human dimension. In this way, we could further control for valence in participants' self-descriptions at both times self- and other-described humanness was measured.

Method

Participants

Based on the smallest effect size that we detected in Study 1 for self-dehumanization and re-humanization effects (d=.63), a power analysis using PANGEA (Westfall, 2016) showed that a sample of 23 participants in each condition (total N = 92) would be sufficient to detect the same effects with a power of .90. On the basis of this analysis, we stopped sampling participants when 221 people responded to the online questionnaire knowing that only 169 (149 women) responded to the full questionnaire. Their age ranged from 18 to 53 (M=23.37, SD=4.28). All were Italian native speakers.

Procedure

In the current study, we compared participants responses in four "enemy" conditions. While the "full enemy" condition was identical to the one presented in Study 1, three extra enemy conditions were created. In the "Enemy no aggression" condition participants engaged with the entire procedure except for the request to

imagine themselves aggressing the target person. In the "Enemy no help" condition we eliminated the request to imagine helping the target person. Finally, in the "Enemy no aggression/no help" condition self and other judgments were only taken at time 1 to avoid that these judgments were made immediately one after the other. Apart from these changes the order in which the various measures appeared was the same as in Study 1.

In Study 2, participants had to respond to more specific questions to describe the enemy. Specifically, they were asked to describe the kind of relationship "you have (had) with this person", "what circumstances have brought you to be in conflict with this person", "how long have you known the person", "how did you meet", "do you still see the person", and "what type of emotions do you feel when you encounter the person". After responding to these questions, participants were asked to judge the intensity of the conflict on a 7-point Likert scale (1=*Not at all intense* – 7=*Very intense*).

Half of the sample was then asked to imagine aggressing against the target person in four hypothetical situations. Three of these were identical to the ones used in Study 1, but the cold water experiment was changed with a situation in which participants could indicate how much envious anger they would feel if something good would happen to the target person. The trait humanness measure that comprised 20 traits would follow. Ten traits referred to the human nature dimension (5 traits were high on HN: Emotional, Warm, Open, Receptive, Sensible and 5 traits were low on HN: Cold, Passive, Rigid, Superficial, Listless), while the remaining ten traits were clearly defined on the human uniqueness dimension (5 traits were high on HU: Civil, Educated, Mature, Refined, Rational and 5 traits were low on HU: Childish, Irrational, Uneducated, Lack of self-control, Coarse). Participants had to rate these traits on the extent to which they described the self first and then the target person thinking back at how they imagined behaving in the four hypothetical situations in the full enemy and in the no help enemy conditions. Instead, participants in the other two conditions where they did not have to imagine aggressing the target person, had to rate the self and the target person in the current moment. Participants responded on a 9-point Likert scale (1=*not at all* to 9=*very much*). In addition, three traits were added to the self-judgments. Based on the work of Leach, Ellemers, and Barreto (2007), participants' moral engagement was measured by asking them to judge themselves on

how honest, trustworthy, and sincere they felt. In order to control for valence at the trait level, we added 9 traits that were selected on the basis of a pretest. All 9 traits were not well defined on the human dimension; their mean ranging from 4.06 to 4.28 on a 7-point scale on both human nature and human uniqueness judgments. At the same time these traits were clearly positively (M=5.79 on a 7-point scale) or negatively valenced (M=2.87 on a 7-point scale). Specifically, the selected traits were: organized, efficient, serene, concrete, relaxed, thin-skinned, dull, irascible, and impatient. Furthermore, all participants were requested to respond to the PANAS (Watson, Clark, & Tellegen, 1988) to measure their mood, after responding to the self-esteem measure. Mood and self-esteem were only measured immediately after the first self- and other humanness judgments.

Results and discussion

After verifying that the aggression, helping, self-esteem, mood and the intensity of the conflict did not differ between the four conditions, we used a similar two step analytic strategy as in Study 1 (See SOM for a full description of the analysis). In the first step, we only considered participants' humanness judgments at time 1, in order to verify whether participants only self-dehumanized when they had the chance to treat the enemy immorally. Pairwise comparisons indicated that participants who imagined behaving aggressively towards the "enemy" judged the self significantly less human compared to those who just had to think of a conflictual other, $F(1,165)=39.29$, $p<.001$, $\eta^2_p=.19$. There was a similar finding for perceptions of the enemy, however, the effect was marginally significant, $F(1,165)=3.66$, $p=.06$, $\eta^2_p=.02$ (see Figure 2). Confirming Hp1, these results provide evidence for the idea that people only self-dehumanize when they engage in immoral conduct. In line with Hp2, in all conditions the conflictual other was seen as significantly less human than the self (all p's<.001). Importantly, all these effects remained significant when self-esteem, mood or the attribution of positive and negative traits to the self and the other were controlled for showing that mood or valence effects do not account for these findings.

In the second step, we focused on the interaction between target and judgments over time separately for the three conditions (in the fourth condition self and other humanness judgments were only measured at Time 1). Note that all these analyses remain unvaried when the attribution of positive and negative traits to the

self and the other were controlled for. In the full enemy condition, the results of Study 1 were replicated (see Figure 2). In line with the tethered humanity hypothesis, both self and other were judged more human at time 2 compared to the first measurement confirming a re-humanization effect of both the self and the other, $F(1,37)=16.74$, $p<.001$, $\eta_p^2=.31$.

In the enemy no aggression condition in which no self-dehumanization effect was observed no changes in self and other judgments occurred over time ($p>.65$). Instead, the tethered humanity effect emerged significantly in the enemy no help condition. Again, both targets were humanized more at time 2 compared to time 1, $F(1,49)=55.48$, $p<.001$, $\eta_p^2=.53$. This effect compared to those of the other conditions reveals that self-dehumanization itself engages attempts at reclaiming one's own humanity and that of the other. In line with Hp3a, when people imagined behaving aggressively towards an 'enemy' and self-dehumanized as a result, they reported an increase in their own humanity and that of the other at time 2. Instead, without self-dehumanization no increase in humanness occurred. Furthermore, and in line with Hp3b, only imagining helping the enemy (without self-dehumanization) does not significantly influence participants' tendency to re-humanize the self and the other. This result suggests as in Study 1 that perpetrators who self-dehumanize recuperate their humanity and re-humanize the victim they originally wronged.

Moral engagement

Similar results to the ones we reported for self-dehumanization, were observed on moral engagement. Whenever people had to imagine aggressing against the enemy, they attributed less moral traits to themselves ($M=6.14$, $SD=2.10$) compared to when they did not ($M=7.34$, $SD=1.33$, $F(1,165)=20.05$, $p<.001$, $\eta_p^2=.11$). In order to test whether participants' sense of moral engagement mediated their tendency to self-dehumanize, a mediation analysis was run following the PROCESS protocol (Model 4) proposed by Hayes (2013). A bias-corrected and accelerated bootstrapping for indirect effects revealed that the indirect effect was significant, as indicated by a 95% confidence interval that did not include 0, based on 5000 bootstrap samples (-.69 – -.26). This analysis provides correlational evidence for Hp4 that participants' level of moral engagement at least partially mediates the relation between behaving immorally and the tendency to self-dehumanize (see Figure 3).

Study 3

In the former studies, we have shown that people who imagine behaving immorally and self-dehumanize respond in a way which suggests they are motivated to restore the humanity of both the self and their victim. This, of course, raises the question of whether the responses of one's victim may play any role at all in this process. In Study 3, we aimed to show that when one's victim rejects attempts at reparation, this interrupts the process of re-humanization of both the self and the other. The tethered humanity hypothesis states that perpetrators who self-dehumanize recuperate their humanity re-humanizing the victim they have originally harmed. Manipulating the attempt at reparation of the victim, allowed us to demonstrate that when the victim cannot be rehumanized because he or she interrupts the reparation process, the self cannot regain its full human status. As such, we were able to show experimentally that the gain in humanness for the self only occurs when people can re-humanize the other they have originally wronged.

In order to test this hypothesis, we created a conflict situation in the lab asking people to conduct an online conversation with another alleged participant. Three conditions were created: a control condition without specific instructions about the tone of the conversation, and two experimental conditions in which participants were instructed to be explicitly rude towards their companion. In the latter two conditions they were instructed to write an apology for their rude conduct and the alleged participant, would either accept (apology accept condition) or reject the apology (apology reject condition). As in the former studies, we expected that only when harm is inflicted, that is when participants were instructed to be rude, they will be inclined to self-dehumanize (Hp1). Given that the conflict was created in the lab and it is hard and somewhat unethical to truly induce negative feelings towards another, we did not expect to observe a clear dehumanization effect of the victims (see Bastian et al., 2013, for similar findings). Finally, we only expected to see the successful re-humanization of the self in the apology accepted condition (Hp3). We reasoned that if a victim does not accept one's apologies, this indicates that the relationship between the perpetrator and the victim remains damaged, inhibiting the ability to re-humanize both the victim and the self.

Method

Participants

Given that the effect sizes for self- and re-humanization effects tended to be stronger in Study 2 compared to Study 1, we based our sample size on the same power analysis we conducted before. For the current design, this analysis showed that a sample of 23 participants in each condition (total $N = 69$) would be sufficient to detect the differences in self-dehumanization and re-humanization effects between conditions with a power of .90. A total of 97 participants (69 women) were enrolled in the current study. Their age ranged from 17 to 24 (M_{age}=18.94, SD=1.47). Of the sample, 68% were born in Australia and 71.1% had English as a first language. All participants were capable of understanding the instructions and responding to the questionnaires. Eight participants were excluded because they did not follow instructions. A further 11 participants were excluded because their responses were not seen as rude enough by the confederate. This left us with a sample of 78 participants (54 women): 32 in the control condition, 26 in the apology accepted condition and 20 in the apology rejected condition.

Procedure

Participants were told that the study explored how people respond to online interactions, and that they would have an online conversation with a participant in another room. For both apology conditions, we used an "illusion of choice" procedure at the beginning of the study. In these conditions, the experimenter stated that most participants had chosen to refrain from conducting a rude conversation with the other participant. Therefore, it would be great if participants would agree to be in this condition, but of course they could choose which experimental condition they wanted to be in. In the control condition, the reverse explanation was provided. Consistent with past work (see Ciarocco, Sommer, & Baumeister, 2001), we expected that by having participants think that they had chosen their condition they would feel a sense of responsibility for what they did during the experiment. Participants who agreed to have a rude conversation were later assigned randomly to the apology accepted or apology rejected condition.

Participants entered a fully enclosed computer room to begin the interaction using *mIRC* internet relay chat software. We told participants that there was another participant who had already arrived and was sitting in a nearby room (this other participant was a confederate who was aware of the experimental condition). After explaining how the chat room software worked, we instructed participants to start by

asking general questions and if they were struggling to find topics to discuss, we suggested them to ask the other person what they were studying.

In the control condition, participants were told to keep the conversation going until the experimenter returned. In the experimental conditions, participants were told "Keep the conversation going, but after about a minute we want you to start being rude to the other participant. Don't say anything deeply offensive; just say something unpleasant, dismissive, or abrupt that makes it clear to the person that you're being a bit nasty to them." If participants could not think of anything rude to say, the experimenter would suggest to say something disparaging about the person's course of study (e.g., "That's a really boring course", "You'll never get a job with that degree"). If the participant had not said anything rude after three minutes, the experimenter returned to the participant's room and asked "How is the conversation going? Has there been any rude interaction yet?" The experimenter told participants "I'll give you another few minutes. Do your best and try to think of something rude to say." If the participant still had not said anything rude after another three minutes, the experimenter moved them on to the next part of the study.

When the participant had said something rude the confederate replied in a way that showed they were offended (e.g., "That's a bit harsh", "I think that's unfair") without retaliating or saying anything that could make the participant feel justified in being rude. After two minutes the experimenter returned to the participant's room and asked them to end the conversation. Next, participants completed the same 20 trait humanness measure that was used in Study 2 (rated on a 5-point scale: 1=*very slightly*, 5=*extremely*). They also completed the PANAS and the self-esteem scale.

After participants finished the questionnaires, the experimenter told those in the control condition to initiate another conversation with the participant and to continue talking in the same way as before. Instead, in the experimental conditions, participants were told to initiate another conversation apologizing for what they said before. The experimenter told participants not to mention that they were asked to be rude as part of the study.

In the "apology accepted" condition, the confederate acknowledged that they had been offended by the participant's comments but accepted the apology. In the "apology rejected" condition the confederate indicated that they had been offended by the participant's comments and did not accept the apology.

The experimenter returned after two minutes and asked the participant to end the conversation again. Next, participants completed the same questionnaires as before (humanness measure, PANAS and self-esteem) answering according to how they were feeling in the present moment. Also, two final questions were added: one asked to evaluate the interaction they had had with the other participant (i.e., How do you think your interaction with the other participant went overall?), the second question asked to evaluate the end of the interaction (How do you think your interaction with the other participant ended?). Both questions were responded to on an 11-point scale (1=*Very unpleasant* to 11=*Very pleasant*)

At the end of the study, the confederate rated the rudeness of the interaction (0=*not rude at all*, 10=*extremely rude*) and how genuine the participant's apology seemed (0=*not genuine at all*, 10=*extremely genuine*). Only participants whose rudeness was judged 6 or higher were retained in the final analyses.

Results and discussion

To analyze participants self and other humanness judgments, we used the same two-step analytical strategy as in the former studies (see SOM for a full description of the analyses). Looking only at the humanness judgments at time 1 revealed that the other is always perceived as more human than the self, $F(1,75)=45.11$, $p<.001$, $\eta^2_p=.38$ (see Figure 4). Given that the conflict was created in the laboratory without inducing true hostile feelings towards the victim, no dehumanization of the victim was expected. Nonetheless, and confirming Hp1 there was evidence of a self-dehumanization effect. Pairwise comparisons between conditions for the self and the other separately indicated that in both experimental conditions participants attributed themselves less humanness compared to the control condition, $F(2,75)=12.69$, $p<.001$, $\eta^2_p=.25$. There was a similar tendency to attribute less humanness to the other in the experimental conditions, $F(2,75)=4.11$, $p=.02$, $\eta^2_p=.10$, even though the pairwise comparisons only reached conventional levels of significance comparing the apology reject and the control condition ($p=.023$).

The second type of analysis focused on the interaction between target and judgments over time separately for the three conditions. In the Control condition, self and other judgments did not change over time ($p>.22$). In the apology accepted condition, the gain of humanness over time was clearly present for both targets, but it tended to be stronger for the self, $F(1,25)=45.99$, $p<.001$, $\eta^2_p=.65$, compared to the

other, $F(1,25)=17.49$, $p<.001$, $\eta^2_p=.41$. This result provided support to the tethered humanity effect; that when the apology was accepted the humanness of both the self and the target was reinforced.

Finally, in the apology rejected condition, where the victim refuses the participant's apology, no re-humanization occurred. While the other significantly lost humanness over time, $F(1,19)=24.39$, $p<.001$, $\eta^2_p=.56$, the gain of humanness for the self was only marginally significant, $F(1,19)=3.56$, $p=.08$, $\eta^2_p=.16$. Importantly, comparing the gain of humanness for the self over time (i.e., humanness attributed to the self at Time 2 minus the humanness attributed to the self at Time 1) between all conditions, the apology accepted condition was significantly different from the other two, $F(2,75)=11.24$, $p<.003$, $\eta^2_p=.23$, showing that especially in this condition the humanness attributed to the self increased significantly over time ($p<.001$ and $p=.05$, comparing the control and apology reject condition respectively). Taken together these findings provide support for Hp3 that when the apology of perpetrators gets accepted, they are more likely to be able to re-humanize themselves, because they can reinforce the human status of their victims. When their attempt for reconciliation is interrupted, instead, we found a tendency for perpetrators to dehumanize their victims not allowing them to rehumanize the self.

Study 4

The former studies analyzed the tethered humanity hypothesis from the perspective of the perpetrator. In Study 4, we focused on the victim's perspective creating an experiment in which participants played the Cyberball game (Williams, 2000), a procedure that has been employed in more than 200 published studies to manipulate social ostracism in the laboratory (Hartgerink, van Beest, Wicherts, Williams, Peters, & Ratliff, 2015). In this game, participants play a virtual ball tossing game with two ostensible other participants. Three conditions were created: an inclusion condition in which participants were fully included in the game; and two exclusion conditions in which participants were actively excluded from the game by the other two participants. In one exclusion condition, at the end of the game one of the players apologized for his/her conduct (exclusion with apology condition), while in the other exclusion condition the other player explicitly did not apologize (exclusion without apology condition). As in the former studies, self and other humanness was measured both immediately after playing the game and again at the

end of the experiment when apologies were received or not. In line with previous research (Bastian & Haslam, 2010), we expected the victims to self-dehumanize (Hp1) and dehumanize (Hp2) the other players when they were excluded rather than included in the game. Because of their tendency to self-dehumanize, we expected victims to be motivated to regain their lost humanity. Following the tethered humanity hypothesis, one way to do this is re-humanizing the person that has originally wronged them. Specifically, one can expect that sincere remorse on the part of the perpetrator might lead victims to accept the perpetrator's apology, forgive him or her and rehumanize both the self and the perpetrator as a result. On the contrary, when a perpetrator refrains from expressing a sincere apology, victims might not be able to regain a full human status; the perpetrator shows no regret, does not acknowledge his or her wrong doing and can therefore not be rehumanized. The interruption of the re-humanization process of the other may therefore impede the possibility to rehumanize the self. Therefore, we expected the re-humanization of both the self and the other only in the exclusion with apology condition, but not in the exclusion condition were no apology was granted (Hp3).

Method

Participants

We followed the same recruitment strategy as in Study 3. A total of 104 participants (92 women) were enrolled in the current study. Their age ranged from 20 to 42 (M_{age}=21.07, SD=0.52). Apart from 3 participants, all were Italian native speakers. Still, the linguistic capacities of all participants were sufficient to understand the instructions and respond to the questionnaires.

Procedure

Participants were accommodated in front of the computer and informed that they would participate in an experiment on "mental visualization" while playing an interactive game online with two other participants that were allegedly seated in other laboratories. After participants were informed about the full procedure of the experiment and gave their informed consent, the female experimenter would leave the laboratory briefly, ostensibly to check whether the other participants were ready to start the experiment. Once she came back, she informed participants that the other players were almost ready and that they could press the start button. After they

reported their own initials, all participants were shown a screen for 5 sec. that stated "Waiting for other participants …". These efforts helped to increase the credibility of the story that there were two other participants also playing the game.

Immediately afterwards participants started to play the Cyberball game. Each player was represented with a small animated figure that was identifiable with the player's initials (the other players were always called "MD" on the right and "DL" on the left). Participants were instructed to choose freely which player they wished to throw the ball to by mouse-clicking on the other players initials whenever their animated figure would catch the ball. The game was programmed to last for about 50 trials and lasted approximately two and a half minutes. In the inclusion condition, participants received the ball after every other throw from each of the other two ostensible players meaning that they received the ball a fair one third of the time. In the exclusion conditions, instead, participants received the ball once from each player at the beginning of the game, and then never received it again afterwards.

After the game, participants were asked to judge how they perceived themselves and both the other players during the game on 12 humanness traits that were selected from the 20 traits used in Study 2 and 3 (Open, Emotional, Sensible, Superficial, Rigid, and Cold for the HN traits and Rational, Refined, Civil, Childish, Lack of self-control and Irrational for the HU traits). All judgements were made on a 7-point scale (1=*not at all* to 7=*very much*). They also completed the PANAS and the self-esteem scale.

After participants finished the questionnaires, they were given the possibility to send a message to one of the players (always the player that appeared on the right side of the screen), while they would receive a message from the player that appeared on the left side of the screen. After they sent the message, a screen would appear for approximately 5 sec. that stated "Sending message …", after which they could visualize the message they received from the other player. In the inclusion condition, the message always read: "Hello. The game seemed a bit monotonous but it was still an interesting experience, because I have never participated in an interactive experiment before. I hope it was like that for you too." In the exclusion condition with an apology, participants read: "Hello. I realize that I excluded you during the game ... I guess that must have been a bad experience for you and I apologize. I hope you're not mad at me." In this message, the apology was not only clearly stated, but the

player also acknowledged the suffering of the victim. Finally, in the exclusion condition without apology, participants received the following message "Hello. The game seemed pretty repetitive, but it didn't last too long. I guess you got bored because we never considered you, but that doesn't matter ...it was fine for me". This message expresses the player's awareness that the participant must have felt bad during the game, but that he or she did not care about this.

Next, participants responded to an adaptation of the Transgression-Related Interpersonal Motivations questionnaire (TRIM, McCullough, Fincham, & Tsang, 2003) to measure participants willingness to forgive the player they received a message from. The TRIM is made up of three sub-scales: Avoidance Motivations, Benevolence Motivations and Revenge Motivations. Specifically, the revenge items were excluded as they were deemed too harsh in the current situation, three benevolence (e.g., I forgive him/her for what he/she did to me) and five avoidance items (e.g., I keep as much distance between us as possible) were retained and slightly adapted to fit the current situation. Participants responded on a 5-point scale (1=*completely disagree* to 5=*fully agree*). Afterwards, participants completed the same questionnaire as before judging how they described themselves and the other players in human terms according to how they were feeling in the present moment.

In the end, as in Study 3, participants responded to two final questions on a 7-point scale (1=*very unpleasant* to 7=*very pleasant*) indicating how their interaction with the other players went overall and how they thought their interaction with the other participants ended. The experiment took about 20 minutes to complete and participants were debriefed and informed that the other players were not actually present and that the Cyberball game followed a pre-programmed script.

Results and discussion

Forgiveness

Comparing participants willingness to forgive as a function of condition indicated significantly less willingness to forgive in the exclusion without apology condition (M=3.31, SD=.90) compared to both the exclusion with apology (M=3.94, SD=.73, p=.004) and inclusion conditions (M=3.86, SD=.69, p=.013), which did not differ (p=.97). Critically, the significant difference between both exclusion conditions indicated that the apology of player 1 allowed participants to forgive his or her immoral conduct.

Self- and other-humanness

The attribution of humanness to the self and to the other two players was analyzed using the same two-step analytic strategy as in the former studies. In a first step we only considered participants' humanness judgments at time 1 to verify whether participants self-dehumanized when becoming a victim of social ostracism and dehumanized the players that excluded them. In line with Hp1, pairwise comparisons indicated that the self was humanized significantly less in both exclusion conditions relative to the inclusion condition, $F(2,101)=12.05$, $p<.001$, $\eta^2_p=.19$, while confirming Hp2, the other two players were significantly dehumanized when they excluded the participant from the game ($F(2,101)=32.02$, $p<.001$, $\eta^2_p=.39$ and $F(2,101)=27.67$, $p<.001$, $\eta^2_p=.35$ for player 1 and 2 respectively, see Figure 5).

In a second step, testing the re-humanization effect (Hp3), we analyzed the interaction between the self and other humanness judgments at both times of measurement in each condition separately. In the inclusion condition, all targets were judged as more human by the end of the experiment compared to immediately after playing the game, $F(1,68)=7.37$, $p=.01$, $\eta^2_p=.18$. In the exclusion with apology condition, pairwise comparisons indicated that all targets regained humanness over time, but this effect was especially significant for the self and player 1 who explicitly apologized for his behavior during the Cyberball game (both p's$<.001$). Also, the second player who did not apologize was seen as somewhat more human at time 2 compared to time 1 ($p=.01$), but he or she was judged significantly less human than both the self and player 1 who granted the apology (both p's$<.001$). Moreover, at time 2 after the apology the self and player 1 were judged as equally human ($p=.88$).

Finally, in the exclusion without apology condition, contrary to our expectations, pairwise comparisons indicated that only the self was significantly judged as more human at time 2, even though no apology was granted ($p=.001$). Importantly, the other players were seen as equally less human than the self at both times of measurement. To further understand this unexpected finding, we further analyzed participants responses on the forgiveness scale in this condition and noticed that while some participants were clearly unwilling to forgive the player who did not apologize, others did. Therefore, we split the participants in this condition on the median of the forgiveness scale ($M_e=3.22$) allowing us to compare those who did forgive the other players from those who did not, even when no apology was given. A

closer look at Figure 6 shows that neither the self nor the other players gained humanness over time when participants did not forgive the other players (all *p*'s>.28), while when they granted forgiveness to the other players even when no apology was given both the self and the other players were seen as significantly more human at time 2 compared to time 1 (all *p*'s<.03). As such, this finding can be interpreted in line with our hypothesis. Even though player 1 did not apologize for his or her immoral behavior, only those participants who decided to forgive the person managed to both rehumanize the others and themselves.

General discussion

People cannot exist independently of their relationships with others. This also means that our humanity is caught up in theirs. When we treat others with dignity and respect, we perceive them as fully human and feel fully human ourselves. We can, however, tarnish our own human self-concept and that of others as well, when we are the perpetrators or victims of interpersonal harm and immoral acts. Starting from this interdependent perspective on dehumanization, across four studies we provided empirical evidence for self-dehumanization and how it relates to the tethered humanity hypothesis. Study 1 demonstrated that thinking about an "enemy" rather than a friend, made people more willing to imagine aggressing against that person and self-dehumanized as a result. Self-dehumanization in turn motivated participants to regain a full human status and re-humanize the "enemy". In Study 2, some of the processes that play a role in the initial dehumanization of the self, the dehumanization of the other, and furthermore the tethered humanity hypothesis, were identified. As far as the former is concerned, self and other dehumanization was triggered by different mechanisms. While feelings of hostility were sufficient to dehumanize the other, it was engaging in immoral behavior that triggered self-dehumanization. Moreover, mediational analyses in this study showed that people will self-dehumanize to the extent that they morally engage with their wrongdoings. Self-dehumanization, in turn, enables the tethered humanity effect to occur. It frustrates a basic motivation to see oneself as human and therefore motivates people to repair their relationships with others in order to regain a full human status. Interestingly, self-dehumanization motivated the re-humanization of the other even in the absence of any (imagined) help offered to the victim. Study 3 demonstrated that the responses of the victim matter. When the victim did not accept the perpetrators' attempts to reconnect, the re-

humanization process was interrupted preventing perpetrators from regaining their lost humanity, presumably because they could not reinforce the humanity of their victim. Finally, in Study 4, the tethered humanity hypothesis was analyzed from the perspective of the victim. Replicating previous research (Bastian & Haslam, 2010), we found victims self-dehumanized because of the treatment they received. Nonetheless, and in line with the tethered humanity hypothesis, we found that when the perpetrator apologized or when victims forgave the perpetrator even when no apology was offered, they managed to re-humanize the perpetrator and therefore the self. Indeed, when no forgiveness was granted, victims did not regain their humanity, because they refrained from re-humanizing their perpetrators.

The self-dehumanization effect and the tethered humanity hypothesis are supported with medium effect sizes even when self-esteem, mood and the valence of the characteristics that were attributed to the self and the other were controlled for. As such, these effects are largely independent from global self-evaluations, simply feeling bad and self- and other-(dis)liking. Moreover, the effects emerged using two different dehumanization measures, both when people reflected upon existing conflictual relationships and in a conflict that was created in the laboratory. Finally, and speaking to the generalizability of the effect across cultural contexts, these results were observed in a mostly US (Study 1), Italian (Study 2, Study 4), and an Australian sample (Study 3). Taken together, these findings provide evidence for the underlying mechanisms that make self-dehumanization a functional process that leads to interpersonal reconnection through the re-humanization of the self and the other when interpersonal harm is inflicted.

From dehumanization to re-humanization

The current set of studies demonstrate how processes of dehumanization and re-humanization of both the self and the other can be related. There is, however, one remaining caveat in this link between dehumanization and re-humanization. Specifically, our results suggest that the dehumanization of the other per se does not trigger self-dehumanization, while the re-humanization of the other enables re-humanization of the self. Indeed, in the former case it seems people need to treat the other immorally and acknowledge their responsibility or become victims of such treatment before they self-dehumanize, while in the latter imagining helping the victim does not appear to make a difference. The re-humanization of the self and the

other seem to co-occur regardless of other actions. A possible explanation for this apparent inconsistency stems from the differential involvement of the self in both processes. The dehumanization of another person does not necessarily imply a strong involvement of the self. Dehumanization is a strategy of moral disengagement that allows keeping the self detached and serves to avoid self-sanctions (Bandura, 1999). Therefore, it is not a surprise that the dehumanization of another person per se does not lead to self-dehumanization. Instead, when people are caught on the spot or asked to reflect on their own wrong-doing, people will more likely engage with their immoral actions and self-dehumanize as a result. The re-humanization process that we demonstrate in the current article, however, always implies a form of self-involvement. To the extent that the re-humanization process is driven by people's motivation to recuperate their own lost humanity, self-motives will always play a central role. The re-humanization of the other, therefore, can become a way to recuperate one's own lost humanity, as long as the other person accepts the attempt to reconnect.

Implications for conflict management and reconciliation

The re-humanization of the other might be an important step in the reconciliation process. Increasing the perceived humanness of a target has been linked with a range of positive and conciliating outcomes such as, empathy (Čehajić, Brown, & Gonzalez, 2009), perspective taking (Vaes, Leyens, & Paladino, 2004), respect (Renger, Mommert, Renger, & Simon, 2016), forgiveness (Tam, Hewstone, Cairns, Tausch, Maio, & Kenworthy, 2007), and helping behavior (Andrighetto, Baldissarri, Lattanzio, Loughnan, & Volpato, 2014; Vaes et al., 2003). The tethered humanity hypothesis and the psychological mechanisms that it unveils could therefore be informative in the way a range of conflicting situations are managed. Dehumanization has been described as a central process in bullying (Pozzoli, Gini, & Vieno, 2012; van Noorden, Haselager, Cillessen, & Bukowski, 2013), intimate partner violence (Bastian, 2019; Pacili, Pagliaro, Loughnan, Gramazio, Spaccatini, & Baldry, 2017) and social ostracism (Bastian & Haslam, 2010). In these contexts, encouraging perpetrators to acknowledge their moral wrongdoings and pushing them to engage morally with their acts, may increase the likelihood that they self-dehumanize and re-humanize their victims as a result. As such, self-dehumanization and its resulting re-

humanization of both the self and the other might be an important step in the reparation of conflicting relationships.

Limitations and Future research

Even though the current findings strongly suggest that perpetrators and victims of interpersonal harm tether their moral self-image to each other, we are neither claiming that this will always happen, nor that it is the most common or natural response. Research has shown that perpetrators often engage in ego-protective responses preferring to justify their own wrongdoing rather than acknowledge it (Bandura, 1999), making it unlikely that they self-dehumanize in the first place. In a similar vein, victims might be prone to seek revenge rather than forgive their culprits (Jackson et al, 2019). In victims of maltreatment, feelings of anger and shame have been shown to accompany their tendency to self-dehumanize (Bastian & Haslam, 2011), both emotions that might motivate victims to seek revenge (Brown, 1970; Chester & DeWall, 2017). To avoid these alternative routes, we made it hard in our experiments for perpetrators to fail to acknowledge their immoral conduct, given that they just reported or demonstrated their (imagined) aggressive or rude behavior and were asked to reflect on it when describing themselves. In addition, victims were confronted with a sincere apology that recognized their suffering, making it easier for them to forgive the other player. Future research might therefore test the tethered humanity hypothesis in natural settings where people can freely choose to apologize or forgive.

From this perspective the tethered humanity hypothesis might be seen as a possible intervention in interpersonal conflict, a desired route that might lead to true reconciliation. For a conflict to lead to reconciliation, it is important that perpetrators morally engage with their wrongdoings and feel sincere remorse, while the victims need to be able to reclaim their humanness, dignity and respect. These concepts are central in Botcharova's (2001) model of reconciliation in which the re-humanization of the other is a central element that precedes forgiveness, justice, and a public apology. Gobodo-Madikizela (2002) emphasizes that genuine remorse implies the acknowledgment of one's wrong doings. Remorse, however, will only lead to forgiveness if a clear attempt is made to make a human connection with the other person. In the expression of remorse, perpetrators are asking to be re-admitted to the

community of humanity, while victims need to feel that their suffering is recognized and their humanness is fully acknowledged.

Taken together, the current work provides evidence, for the first time, that our humanity can become tethered with that of others in animosity as well as in benevolence. It is our hope that the current work might pave the way to create new reconciliation strategies and approaches for conflict management.

References

Andrighetto, L., Baldissarri, C., Lattanzio, S., Loughnan, S., & Volpato, C. (2014). Human-itarian aid? Two forms of dehumanization and willingness to help after natural disasters. *British Journal of Social Psychology, 53*, 573-584. doi:10.1111/bjso.12066

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

Bastian, B. (2019). A dehumanization perspective on dependence in low-satisfaction (abusive) relationships. *Journal of Social Personal Relationships, 36*, 1421-1440. doi: 10.1177/0265407519835978

Bastian, B., Denson, T.F., Haslam, N. (2013). The Roles of Dehumanization and Moral Outrage in Retributive Justice. *PLoS ONE 8(4)*: e61842. doi:10.1371/journal.pone.0061842

Bastian, B., & Haslam, N. (2010). Excluded from humanity: The dehumanizing effects of social ostracism. *Journal of Experimental Social Psychology*, *46*, 107-113. doi:10.1016/j.jesp.2009.06.022

Bastian, B., & Haslam, N. (2011). Experiencing dehumanization: Cognitive and emotional effects of everyday dehumanization. *Basic and Applied Social Psychology*, *33*, 295-303. doi:10.1080/01973533.2011.614132

Bastian, B., Jetten, J., Chen, H., Radke, H., Harding, J.F., & Fasoli, F. (2013). Losing our humanity: The self-dehumanizing consequences of social ostracism. *Personality and Social Psychology Bulletin, 39*, 156-169. doi:10.1177/0146167212471205

Bastian, B., Jetten, J., & Haslam, N. (2014). An interpersonal perspective on dehumanization. In P.G., Bain, J. Vaes, & JPh Leyens (Eds.) *Humanness and*

dehumanization (pp. 205-224). New York: Psychology Press. doi:10.4324/9780203110539

Bastian, B., Jetten, J., & Radke, H. (2012). Cyber-dehumanization: Violent video game play diminishes our humanity. *Journal of Experimental Social Psychology*, *48*, 486-491. doi:10.1016/j.jesp.2011.10.009

Bastian, B., Laham, S.M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. *British Journal of Social Psychology*, *50*, 469-483. doi:10.1348/014466610X521383

Bastian, B., & Loughnan, S. (2017). Resolving the meat-paradox: A motivational account of morally troublesome behavior and its maintenance. *Personality and Psychology Review, 21*, 278-299. doi:10.1177/1088868316647562

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*, 994-1005. doi: 10.1037//0022-3514.59.5.994

Botcharova, O. (2001). Implementation of track two diplomacy: Developing a model of forgiveness. In G. Raymond, S.J. Helmick, & R.L. Petersen (Eds.), *Forgiveness and reconciliation: Religion, public policy, and conflict transformation* (pp. 279– 305). Philadelphia: Templeton Press.

Brandt, M. J., & Reyna, C. (2011). The chain of being: A hierarchy of morality. *Perspectives on Psychological Science*, *6*, 428-446. doi:10.1177/1745691611414587

Brown, B.R. (1970). Face-saving following experimentally induced embarrassment. *Journal of Experimental Social Psychology, 6(3)*, 255–271. doi: 10.1016/0022-1031(70)90061-2

Buckels, E.E., & Trapnell, P.D. (2013). Disgust facilitates outgroup dehumanization. *Group Processes and Intergroup Relations, 16*, 771-780. doi:10.1177/1368430212471738

Čehajić, S., Brown, R., & Gonzalez, R. (2009). What do I care? Perceived ingroup responsibility and dehumanisation as predictors of empathy felt for the victim group. *Group Processes & Intergroup Relations, 12*, 715–729. doi:10.1177/1368430209347727

Chester, D.S., & DeWall, C.N. (2017). Combating the sting of rejection with the pleasure of revenge: A new look at how emotion shapes aggression. *Journal of Personality and Social Psychology, 112(3)*, 413–430. doi: 10.1037/pspi0000080

Ciarocco, N.J., Sommer, K.L., & Baumeister, R.F. (2001). Ostracism and ego depletion: The strains of silence. *Personality and Social Psychology Bulletin*, *27*, 1156-1163. doi:10.1177/0146167201279008

Delgado Rodriguez, N., Rodriguez-Pérez, A., Vaes, J., Betancor Rodriguez, V., & Leyens, J.Ph. (2012). Contextual variations of infrahumanization: The role of physical context and territoriality. *Basic and Applied Social Psychology, 34*, 456-466. doi: 10.1080/01973533.2012.712020

Demoulin, S., Cortes, B.P., Viki, T.G., Rodriguez-Perez, A., Rodriguez-Torres, R., Paladino, M.P., et al. (2009). The role of in-group identification in infra-humanisation. *International Journal of Psychology, 44*, 4–11. doi:10.1080/00207590802057654

Fehr R, Gelfand MJ, & Nag M. (2010). The road to forgiveness: a meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin,* 136, 894–914. doi: 10.1037/a0019993

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.

Gobodo-Madikizela, P. (2002). Remorse, forgiveness, and rehumanization: Stories from South Africa. *Journal of Humanistic Psychology, 42*, 7–32. doi:10.1177/0022167802421002

Goldenberg, J.L., Pyszczynski, T., Greenberg, J., Solomon, S., Kluck, B., & Cornwell, R. (2001). I am not an animal: Mortality salience, disgust, and the denial of human creatureliness. *Journal of Experimental Psychology: General*, *130*, 427-435.

Goldenberg, J.L., Heflick, N., Vaes, J., Motyl, M., & Greenberg, J. (2009). Of mice and men and objectified women: A terror management account of infra-humanisation. Group Processes & Intergroup Relations, 12, 763–776. doi:10.1177/1368430209340569

Goldstein, J.H., Davis, R.W., & Herman, D. (1975). Escalation of aggresion: Experimental studies. *Journal of Personality and Social Psychology, 31*, 162-170. doi:10.1037/h0076241

Gray, K. Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23(2)*, 101-124. doi:10.1080/1047840X.2012.651387

Harmon-Jones, E., & Mills, J. (Eds.). (1999). Cognitive dissonance: Progress on a pivotal theory in social psychology. Washington, DC: American Psychological Association.

Harris, L.T., & Fiske, S.T. (2006). Dehumanising the lowest of the low: Neuroimaging responses to extreme outgroups. *Psychological Science*, *17*, 845-853. doi:10.1111/j.1467-9280.2006.01793.x

Hartgerink, C.H.J., van Beest, I., Wicherts, J.M., Williams, K.D. (2015). The Ordinal Effects of Ostracism: A Meta-Analysis of 120 Cyberball Studies. *PLoS ONE 10(5): e0127002*. doi:10.1371/journal. pone.0127002

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*, 252-264. doi:10.1207/s15327957pspr1003_4

Haslam, N., Bain, P., Douge., L., Lee, M., Bastian, B., (2005). More human than you: Attributing humanness to self and others. *Journal of Personality and Social Psychology, 89*, 937-950. doi:10.1037/0022-3514.89.6.937

Hayes, A.F. (2013). *Introduction to mediation, moderation and conditional process analysis, 2nd edition: A regression-based approach*. NY: The Guilford Press.

Jackson, J.C., Choi, V.K., & Gelfand, M.J. (2019). Revenge: A multilevel review and synthesis. *Annual Review of Psychology, 70*, 319-345. doi: 10.1146/annurev-psych-010418-103305

Jordon, J., Mullen, E., & Murnighan, J.K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin*, *37*, 701-713. doi:10.1177/0146167211400208

Kenworthy, J.B., Miller, N., Collins, B.E., Read, S.J., & Earleywine, M. (2011). A trans-paradigm theoretical synthesis of cognitive dissonance theory: Illuminating the nature of discomfort. *European Review of Social Psychology, 22*, 36-113. doi: 10.1080/10463283.2011.580155

Kouchaki, M., Dobson, K.S.H., Waytz, A., & Kteily, N.S. (2018). The link between self-dehumanization and immoral behavior. *Psychological Science, 29*, 1234-1246. doi:10.1177/0956797618760784

Leach, C.W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-

groups. *Journal of Personality and Social Psychology, 93*, 234–249. doi:10.1037/0022-3514.93.2.234

Martens, A., Kosloff, S., Greenberg, J., Landau, M. J., & Schmader, T. (2007). Killing begets killing: Evidence from a bug-killing paradigm that initial killing fuels subsequent killing. *Personality and Social Psychology Bulletin*, *33*, 1251-1264. doi: 10.1177/0146167207303020

McCullough, M.E., Fincham, F.D., & Tsang, J. (2003). Forgiveness, forbearance, and time: The temporal unfolding of Transgression-Related Interpersonal Motivations. *Journal of Personality and Social Psychology, 84*, 540-557. doi: 10.1037//0022-3514.84.3.540

McCullough, M. E., Root, L. M., Tabak, B. A., & Witvliet, C. v. O. (2009). Forgiveness. In S. J. Lopez & C. R. Snyder (Eds.), *Oxford library of psychology. Oxford handbook of positive psychology* (pp. 427-435). New York, NY, US: Oxford University Press.

Ohbuchi, K., Kameda, M., & Agarie, N. (1989). Apology as aggression control: Its role in mediating appraisal and response to harm. *Journal of Personality and Social Psychology*, *56*, 219–227. doi: 10.1037//0022-3514.56.2.219

Pacili, M.G., Pagliaro, S., Loughnan, S., Gramazio, S., Spaccatini, F., & Baldry, A.C. (2017). Sexualization reduces helping intentions towards female victims of intimate partner violence through mediation of moral patiency. *British Journal of Social Psychology, 56,* 293-313. doi:10.1111/bjso.12169

Paladino, M.P., Vaes, J., Castano, E., Demoulin, S., & Leyens, J.P. (2004). Emotional infra-humanisation in intergroup relations: The role of national identification in the attribution of primary and secondary emotions to Italians and Germans. *Current Psychology of Cognition, 22*, 519–536.

Pozzoli, T., Gini, G., & Vieno, A. (2012). Individual and class moral disengagement in bullying among elementary school children. *Aggressive Behavior, 38*, 378–388. doi:10.1002/ab.21442

Renger, D., Mommert, A., Renger, S., & Simon, B. (2016). When less equal is less human: Intragroup (dis)respect and the experience of being human. *The Journal of Social Psychology, 156*, 553-563. doi:10.1080/00224545.2015.1135865

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Sachdeva, S., Iliev, R., & Medin, D. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science*, *20*, 523-528. doi:10.1111/j.1467-9280.2009.02326.x

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, 74-78. doi: 10.1177/0963721417741709

Tam, T., Hewstone, M., Cairns, E., Tausch, N., Maio, G., & Kenworthy, J. (2007). The impact of intergroup emotions on forgiveness in Northern Ireland. *Group processes & Intergroup Relations, 10*, 119–136. doi:10.1177/1368430207071345

Tetlock, P.E., Kristel, O.V., Elson, B., Green, M.C., & Lerner, J.S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*, 853-870. doi:10.1037//0022-3514.78.5.853

Vaes, J., Heflick, N.A., & Goldenberg, J.L. (2010). "We are people": Ingroup humanization as existential defense. *Journal of Personality and Social Psychology*, *98*, 750-760. doi:10.1037/a0017658

Vaes, J., Paladino, M.P., & Leyens, J.P. (2004). Perspective taking in an intergroup context and the use of uniquely human emotions: Drawing an E on your forehead. *International Review of Social Psychology, 17*, 5–26.

Vaes, J., Paladino, M.P., Castelli, L., Leyens, J.P., & Giovanazzi, A. (2003). On the behavioral consequences of infrahumanization: The implicit role of uniquely human emotions in inter-group relations. *Journal of Personality and Social Psychology*, *85*, 1016-1034. doi:10.1037/0022- 3514.85.6.1016

van Noorden, T.H.J., Haselager, G.J.T, Cillessen, A.H.N., & Bukowski, W.M. (2014). Dehumanization in children: The link with moral disengagement in bullying and victimization. *Aggressive Behavior, 40*, 320-328. doi:10.1002/ab.21522

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070. doi: 10.1037//0022-3514.54.6.1063

Westfall, J. (2016). PANGEA: Power analysis for general anova design. *Unpublished manuscript. Available at* http://jakewestfall.org/publications/pangea.pdf.

Williams, K.D., Cheung, C.K., Choi, W. (2000). Cyberostracism: Effects of being ignored over the internet. Journal of Personality and Social Psychology, 79, 748-762. doi: 10.1037//0022-3514.79.5.748

Zhong, C.B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*, 1451-1452. doi:10.1126/science.1130726

Table 1: Mean (and SD) self-esteem and positive and negative mood judgments as a function of condition (Study 3).

		Control	Apology accepted	Apology rejected
Time 1	Positive mood (scale 1-5; α=.86)	2.63 (.64)[a]	2.38 (.65)[a]	2.60 (.82)[a]
	Negative mood (scale 1-5; α=.91)	1.33 (.40)[a]	2.25 (.90)[b]	2.41 (1.03)[b]
	Self-esteem (scale 1-6; α=.87)	4.12 (.77)[a]	3.40 (.96)[b]	3.35 (.92)[b]
Time 2	Positive mood (scale 1-5; α=.90)	2.71 (.63)[a]	2.79 (.73)[a]	2.19 (.73)[b]
	Negative mood (scale 1-5; α=.93)	1.24 (.39)[a]	1.55 (.76)[a]	2.12 (.97)[b]
	Self-esteem (scale 1-6; α=.87)	4.28 (.67)[a]	4.05 (.93)[ab]	3.54 (1.00)[b]

Note. Means with a different superscript are significantly different between conditions ($p < .05$, Bonferroni corrected)

Figure 1: Mean self and other humanness judgments after the hypothetical aggressive and helping situations as a function of condition in Study 1 (error bars represent Standard Errors around the mean).

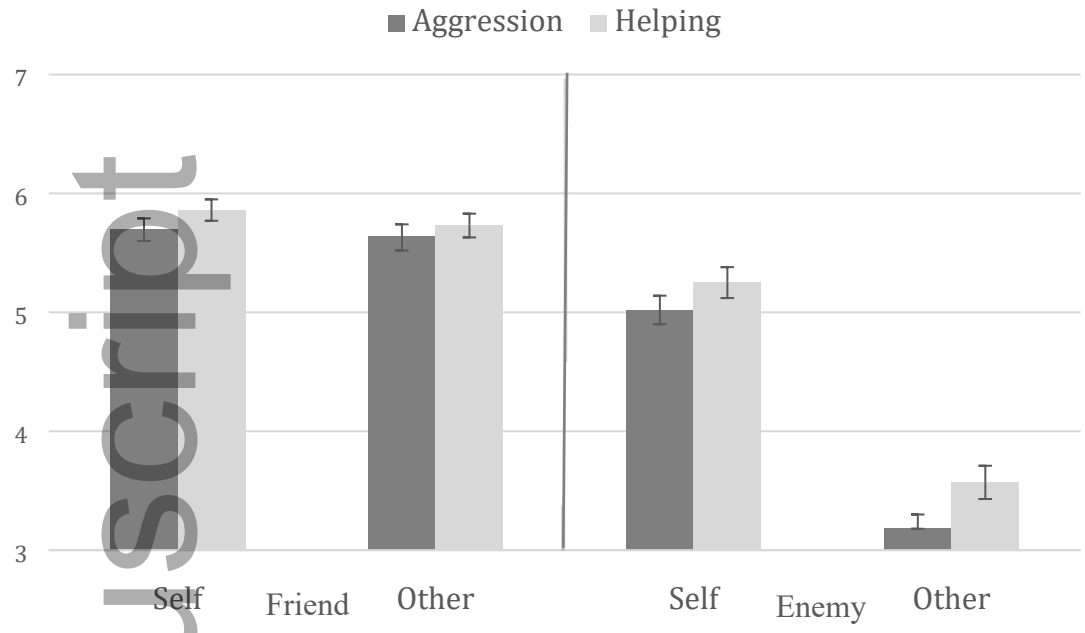

Figure 2: Mean self and other humanness judgments at Time 1 and 2 as a function of condition in Study 2 (error bars represent Standard Errors around the mean).

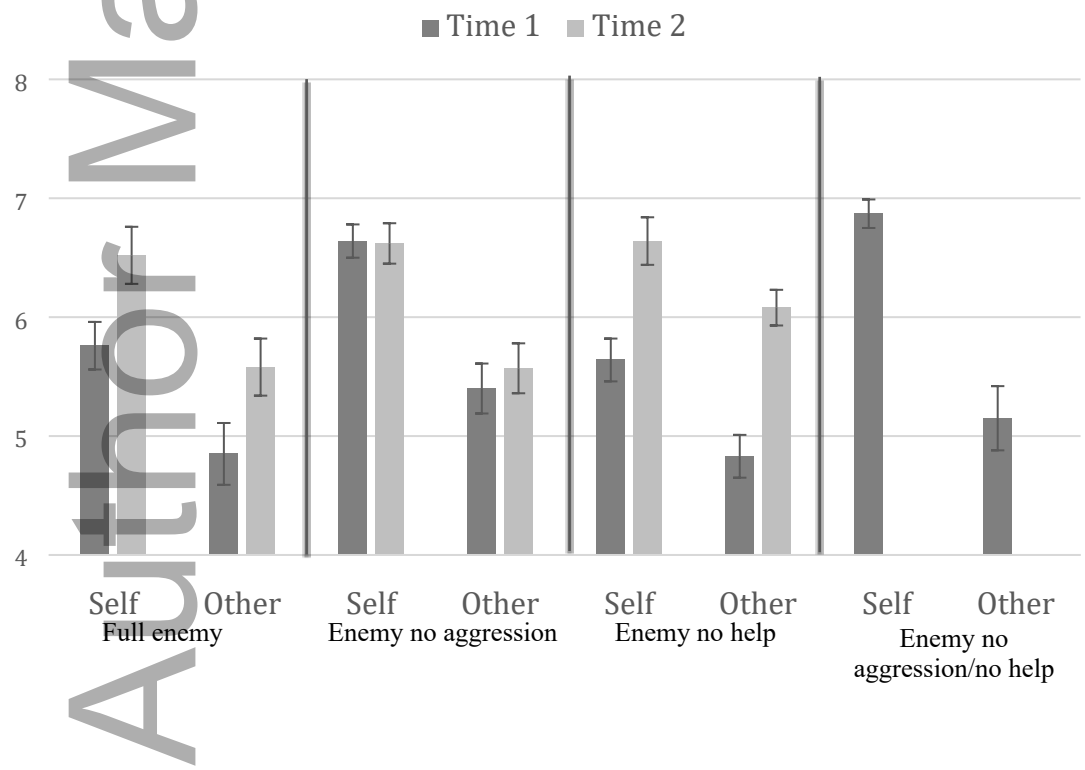

Figure 3: Mediation model testing whether moral engagement mediates the relation between immoral behavior and self-dehumanization (direct effect controlling for the mediator between parentheses).

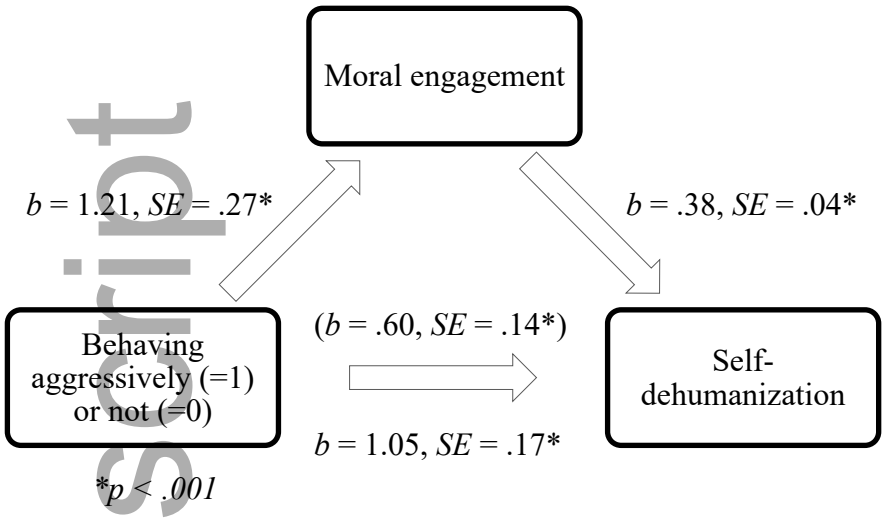

Figure 4: Mean self and other humanness judgments at Time 1 and 2 as a function of condition in Study 3 (error bars represent Standard Errors around the mean).

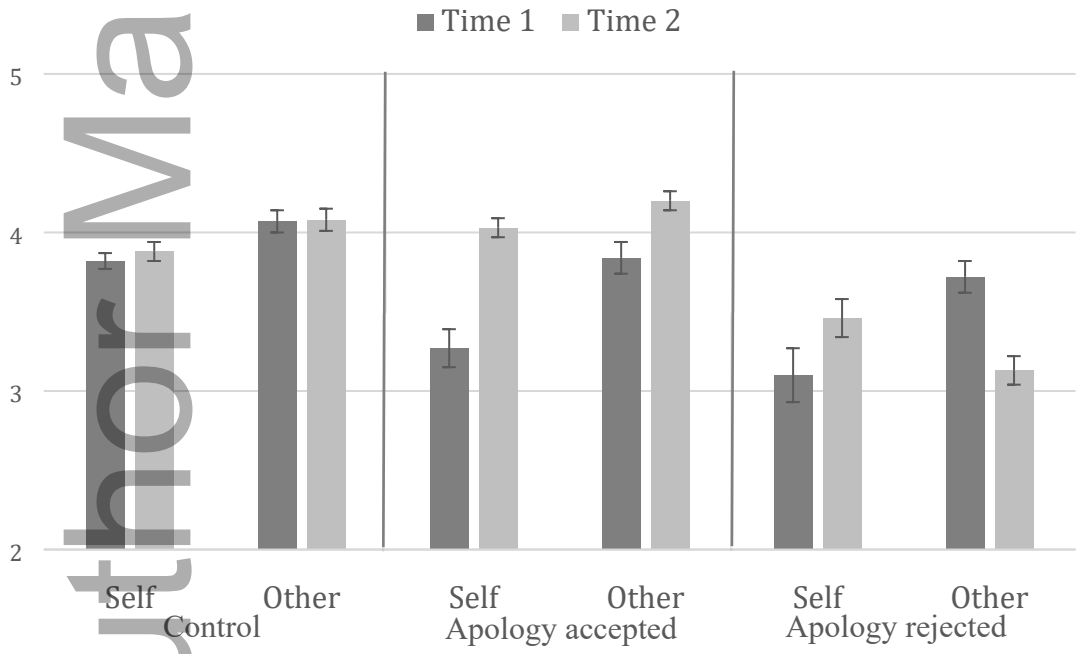

Figure 5: Mean self and other humanness judgments at Time 1 and 2 as a function of condition in Study 4 (error bars represent Standard Errors around the mean).

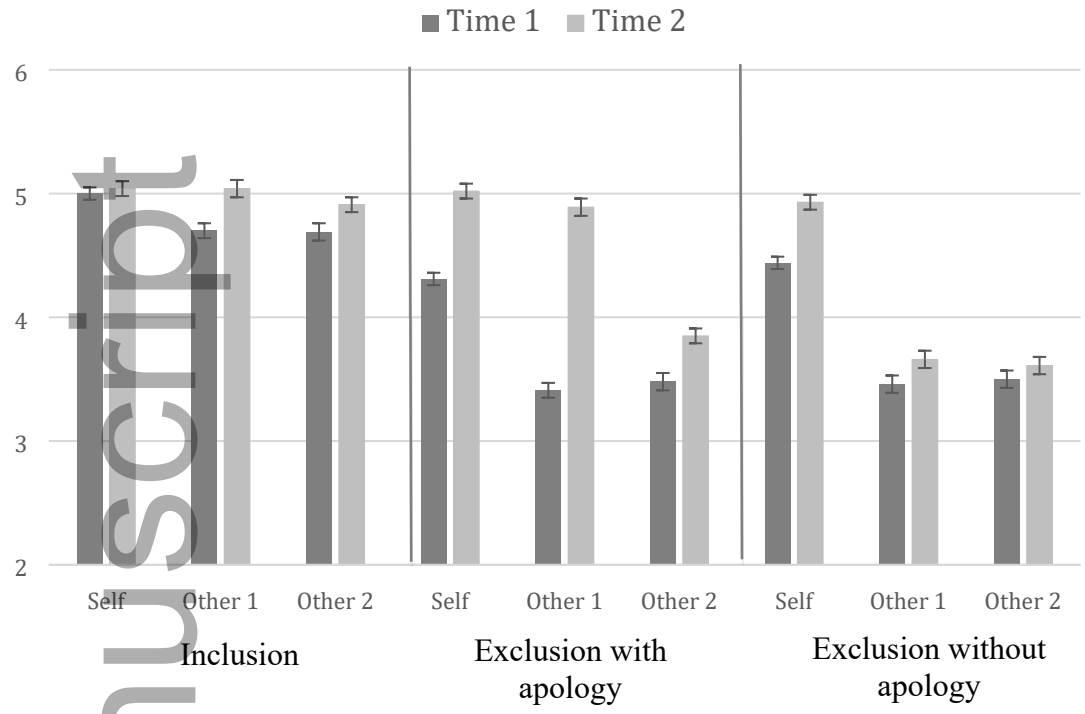

Figure 6: Mean self and other humanness judgments at Time 1 and 2 in the exclusion without apology condition as a function of participants tendency to forgive the perpetrator (error bars represent Standard Errors around the mean).