PsychOpen GOLD

Social Psychological Bulletin

Psychologia Społeczna

Followers Forever: Prior Commitment Predicts Post-Scandal Support of a Social Media Celebrity

Simon Tobias Karg[1], Michelle Lim[2], Simone Schnall[2]

[1] *Department of Political Science, Aarhus University, Aarhus, Denmark.* **[2]** *Department of Psychology, University of Cambridge, Cambridge, United Kingdom.*

Social Psychological Bulletin, 2022, Vol. 17, Article e8283, https://doi.org/10.32872/spb.8283

Received: 2022-02-02 • **Accepted:** 2022-05-26 • **Published (VoR):** 2022-09-06

Handling Editor: Katarzyna Myslinska-Szarek, SWPS University of Social Sciences and Humanities, Warsaw, Poland

Corresponding Author: Simon Tobias Karg, Department of Political Science, Aarhus University, Bartholins Allé 7, DK-8000 Aarhus C, Denmark. E-mail: simonkarg@ps.au.dk

Related: This article is part of the SPB Special Topic "A Festschrift in Honour of Bogdan Wojciszke", Guest Editors: Katarzyna Myślińska-Szarek, Konrad Bocian, & Michał Parzuchowski, Social Psychological Bulletin, 17, https://doi.org/10.32872/spb.v17

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

When learning about wrongdoings of others, people are quick to condemn them and make negative inferences about their character. This tends to not be the case, however, when they hold strong positive feelings toward a transgressor, or consider this person to be part of their ingroup. We investigated the extent to which followers of a social media celebrity, Logan Paul, would still support him after a highly publicized scandal, thus exploring whether they would remain loyal given their prior commitment, or instead, feel especially betrayed and therefore revise their previously positive evaluation of him. Using Distributed Dictionary Representations on a large dataset of YouTube followers (N = 36,464) who commented both before and after the scandal, we found that the more often a person had publicly expressed their approval of the protagonist prior to the scandal, the stronger their post-scandal support was. Similarly, prior engagement was also associated with fewer negative moral emotions, and more positive emotions and attempts to defend the transgressor. Furthermore, compared to non-followers of the celebrity, followers were substantially more supportive of him after the scandal. Thus, highly committed fans failed to update existing moral character evaluations even in light of an extreme moral norm violation, a pattern that is consistent with attempts to reduce cognitive dissonance to maintain a positive evaluation of self and transgressor.

Keywords

person perception, attitudes, commitment, social media, Distributed Dictionary Representation

Highlights

- Holding a person in high esteem biases how new negative information about that person is integrated.
- We investigate how fans of the YouTube celebrity Logan Paul behaved after he engaged in a severe moral transgression by predicting their reactions to Paul's apology video from their pre-scandal comments.
- We find that fans' pre-scandal engagement, as well as their expressed liking towards and social identification with Paul positively predicts post-scandal support.
- This provides field evidence that the more committed people are to positive evaluations of a person, the less likely they will update such evaluations even when facing highly negative, counter-attitudinal information.

Humans are social animals, and living in a society requires the ability to accurately determine who is friend or foe. A key concern therefore is to understand how people form impressions of others' moral character, and furthermore, how they update these impressions when new information, especially of a negative nature, comes to light. To test this question, we examined a high-profile scandal involving a social media celebrity that generated widespread condemnation, to see to what extent devoted followers would subsequently forgive their idol. In other words, we explored how people who are highly committed to someone whom they admire will update their moral character evaluations after this person has engaged in a severe transgression and subsequently apologized for it. We did so by analyzing publicly available YouTube posts related to a high-profile scandal, employing a word vectorization approach that allowed sophisticated extraction of linguistic content from social media data.

The scandal involved the American YouTube celebrity Logan Paul, who since about 2015 has been posting daily videos of entertaining, and mostly harmless pranks on his popular YouTube channel. At the time of the incident, he had about 15 million subscribers, most of them adolescents and young adults. On December 30th 2017, Paul and his entourage visited a Japanese forest that has sadly become known as "suicide forest" because many people have attempted to end their lives there. Shortly after entering the forest, Paul and his team encountered a suicide victim, namely a man who had hanged himself on a tree. Instead of putting the cameras down, however, Paul and his team decided to get a better look at the body. They filmed it close up, and even made inappropriate jokes about the situation, and the deceased. The next day, Paul uploaded the video (with only the victim's face blurred) onto his YouTube channel, and his accompanying remarks

made it clear that he was fully aware of just how shocking the footage would be to watch.

The video spread quickly, garnering around 6.2 million views within a few days, and received widespread media attention (Schimkowitz, 2018). A great deal of public outrage ensued, referred to as the "suicide forest" scandal. Reacting to the outrage, two days after posting the suicide forest video, Paul uploaded an apology video, in which he expressed remorse, apologized to the victim and his fans, and promised to better himself. He then went on a month-long social media hiatus during which he was dropped by advertisers and saw some of his movie arrangements withdrawn. Thus, clearly, Paul's behavior was seen as highly objectionable by many. The scandals therefore offered a real-life context within which to examine how people re-evaluate their moral judgments of a person, especially if they are followers who previously had a highly positive view.

Moral Character: Impression Formation and Updating

Moral character describes a person's inherent moral attributes (Uhlmann et al., 2015; Wojciszke, 2005), and as a construct subsumes a plethora of moral virtues such as being faithful, honest, supportive, kind, understanding, and trustworthy (Narvaez & Lapsley, 2009). Moral character is often assumed to be a stable and essential part of personal identity, defining who a person *really* is (Strohminger & Nichols, 2014). Importantly, when forming moral character evaluations, negative or immoral actions are seen as especially diagnostic (Uhlmann et al., 2015), exemplifying the more general phenomenon of *negativity bias* (Rozin & Royzman, 2001). One reason is that the motives behind moral actions are ambiguous: Actors may merely want to *appear* moral in order to gain benefits, rather than wanting to *be* moral. In contrast, there are no straightforward reasons for wanting to be seen as immoral. Thus, people are highly sensitive to potentially negative character information because it is informative when trying to predict another person's behavior (Cone & Ferguson, 2015).

Applying this discussion to the present context, given that negative character information is weighted more heavily than positive information, it would be plausible that followers of Logan Paul should withdraw their support once they learned about the suicide forest scandal. An alternative possibility, however, is that highly loyal followers are motivated to maintain the positive image of their idol, therefore refusing to take on board incongruent information. This may be due to people either having strong priors about the wrongdoer, or people being willfully ignorant about the negative information (Kim et al., 2020). Several factors influencing this relationship will be reviewed next.

Social Constraints on Character Judgments: Liking and Social Identity

Pre-existing positive feelings toward a person can constrain the degree to which they are held accountable for negative actions. For example, more likeable offenders receive less blame and are more easily forgiven (Bocian et al., 2018). Liking has also been shown to lead to more positive moral character evaluations when it was induced via repeated exposure to a moral actor's face, or facial mimicry of that person, a phenomenon termed the *mere liking effect* (Bocian et al., 2018). Similarly, likeable people are seen as less responsible for negative outcomes of their actions (Nadler & McDonnell, 2012) because these outcomes are seen as less intentional (Newman et al., 2014). Consequently, when judging the transgressions of those who are close to us and whom we like, we grant them a considerable amount of discretion (Forbes & Stellar, 2021). In the context of Logan Paul's suicide forest scandal, his followers already had a highly positive image of him, and therefore might have been more likely to forgive.

This relates to a second factor attenuating the impact of new information on character judgments: group membership or social identity. There is widespread evidence that people use celebrities in how they construct their social identity, and that online forums are ideal places for such processes (Soukup, 2006). Once an idol or leader comes under pressure, people who have idolized this person can perceive threats to this figure as threats to their own self-identity (Branscombe et al., 1999), and will therefore be motivated to defend them if at all possible (Cadinu & Cerchioni, 2001). From the online exchanges in response to Paul's YouTube videos, it is apparent that many of his viewers did indeed create a social identity around the celebrity, referring to themselves as being part of the 'Logang,' (i.e., they devised a label for group membership based on his last name). The outrage caused by the scandal might therefore have resulted in a strong motivational pull for fans to defend Paul's actions to galvanize as a group and protect this part of their social identity.

Having publicly spoken highly of Paul, or otherwise having shown signs of affiliation, followers might have difficulty revising their existing view of him considering the scandal, because their past behavior represented a commitment that cannot easily be taken back. This can be understood using the lens of cognitive dissonance (Festinger, 1962): When behavior conflicts with an attitude, people are highly motivated to bring the two in alignment. Since behavior cannot be altered after the fact, this can only be accomplished by revising the attitude.

With special relevance to the present case, Stoker (1993) analyzed support of Democrats' presidential race candidate Gary Hart, who was having an extra-marital affair. The ensuing scandal's impact on Hart's approval ratings depended on party affiliation, and voters' reasons for endorsing him in the first place: Democrats reacted less negatively to the scandal than Republicans. However, political involvement moderated the scandal's impact within Democrats, such that those who supported Hart because of his policy

program ("issue voters") actually showed *higher* post-scandal approval of Hart than pre-scandal, likely in an effort to reduce cognitive dissonance. Pre-scandal Democrat supporters of Hart who did not place as much importance on Hart's policy, but still preferred him to his two competitors, by contrast, showed a radical negative shift in approval, and they in fact reacted even more negatively to the scandal than most Republicans.

Considering the case of Paul, having shown signs of (public) affiliation might make it more difficult for followers to revise their existing view of him in light of the scandal, because their past behavior represented a commitment that cannot easily be taken back. Consequently, even severe transgressions may not lead to a revocation of support from people who strongly identify with the transgressor.

Indicators of Moral Character Evaluations Following a Transgression

When examining the impact of a scandal on how the wrongdoer is perceived, several factors warrant consideration. People may express approval or disapproval and continued support by generally positive or negative comments. That is, their responses have a broad affective valence. In addition, specific moral emotions are relevant, and indeed, feelings lie at the heart of many evaluations of what is morally right and wrong (Haidt, 2001), and are especially relevant for moral character judgments (Bocian et al., 2018; Giner-Sorolla & Chapman, 2017; Uhlmann et al., 2015). In the current work, we therefore considered the following emotions:

Anger

Anger towards a transgressor of moral norms holds a dominant role in the moral domain, and is especially strong when another individual is harmed, and norms of justice are violated (Graham et al., 2011). Higher levels of anger are expressed when transgressions are more severe, intentional, and non-justifiable (Russell et al., 2013). The angrier a person feels when evaluating a wrong-doing, the more severe is the punishment they deem appropriate (Canton, 2015).

Importantly, anger arises especially when assessing character flaws. For instance, Uhlmann et al. (2014) demonstrated that participants reported greater outrage when explicitly instructed to rate the moral character of bigots before subsequently rating their specific transgressions, compared to a condition with a reversed task order. Thus, immoral acts from persons perceived to have an immoral character lead to stronger condemnation, and the amount of outrage a person expresses can therefore be seen as an indication of how negatively the person evaluates the transgressor's moral character.

Disgust

There is substantial evidence that disgust plays an important role in the context of morality (for reviews, see Giner-Sorolla et al., 2018; Schnall, 2017) and several reasons make disgust especially relevant in the present context. First, the mocking of a suicide victim as in the Logan Paul case constitutes a violation of the purity domain, a key variable in moral considerations (Graham et al., 2011). Second, like anger, disgust is strongly linked to moral character evaluations (Giner-Sorolla & Chapman, 2017), and therefore should be particularly powerful for updating moral character impressions: as a moral emotion concerned with character judgments, disgust leads to withdrawal from transgressors, and attempts to socially isolate them (Giner-Sorolla & Chapman, 2017). Finally, because it is a particularly strong gut reaction, disgust should be especially resistant to the influence of motivated reasoning strategies (Piazza et al., 2013; Russell & Giner-Sorolla, 2011).

Adoration and Compassion

As a result of cognitive dissonance reduction, rather than experiencing the negative emotions usually associated with moral transgressions, highly committed individuals may showcase their continued support by displaying positive emotions reflecting general liking, adoration or compassion towards the transgressor (Thomson & Siegel, 2017). Indeed, these positive sentiments have been linked to forgiveness and continued support after transgressions (Fehr et al., 2010; Forbes & Stellar, 2021).

With particular relevance to the current investigation, a qualitative study of fan responses to celebrity scandals indicated compassion for the transgressor to be a common theme (Finsterwalder et al., 2017). This is also in line with a meta-analysis on forgiveness by Fehr et al. (2010), which demonstrated that perspective taking and trait empathy were positively correlated with forgiveness. A fan's expression of adoration and compassion in response to a scandal can therefore be seen as an indicator of continued support.

The Current Study

Our study investigated the extent to which followers of Logan Paul changed their support after the suicide forest scandal, which was widely considered to be a severe moral violation. To do so, we examined publicly available comments people posted on YouTube in response to Paul's seven videos in the week leading up to the scandal, and compared them to their comments posted to his apology video posted after the scandal.

More specifically, we looked at a commentator's indications of liking, engagement, and social identity before the scandal, and how these related to their general support and specific emotional responses such as anger, disgust, adoration, and compassion, as well as social identity, after the scandal. We expected fans who were highly committed (i.e., followers who had frequently and positively engaged with his videos before the scandal) to remain more supportive of Logan Paul than those who were not, and show more positive (adoration and compassion) and fewer negative emotions (anger and disgust) as

core reactions to Paul's apology. In addition, we hypothesized that expressed pre-scandal liking and social identity would predict post-scandal support and emotions.

In contrast to earlier laboratory studies testing how people update moral character evaluations with hypothetical scenarios, we examined a high-profile scandal that unfolded in real life, and for which we could turn to spontaneous reactions from participants that they shared on a public forum. Our investigation therefore achieved a high level of external validity with naturally occurring responses from tens of thousands of individuals. In addition, we showcase a relatively recent approach to natural language processing, called *Distributed Dictionary Representation* (DDR; Garten et al., 2018), which we used to analyze the YouTube comments regarding relevant linguistic dimensions. This technique enables researchers to tackle research questions using advanced text analysis, as we describe next.

The Distributed Dictionary Representation Method

Distributed dictionary representation (DDR) is a technique for natural language processing (Garten et al., 2018) that builds upon so-called *word vector models*, which organize words according to their semantic meaning in a multi-dimensional vector space (Mikolov et al., 2013). Words that are related in meaning are close to each other in that space. By taking a very large corpus of text – for instance all of Wikipedia – word vectorization algorithms can calculate the semantic distance between a wide range of words.

Conceptually, having such a map of the semantic space makes it easy to compare any two points (e.g., words or sentences) to each other by measuring the distance between them. DDR uses this principle to compare texts to a specific concept of interest. To do so, DDR relies on *concept dictionaries*, which are lists of words associated with a given concept (e.g., love, forgiveness, etc.). By measuring the distance in vector space between a given text (e.g., a YouTube comment) to a pre-defined concept dictionary, one can measure the degree to which the text is related to a concept. Importantly, concept dictionaries can be relatively short because the words in these dictionaries are not counted, as in traditional approaches (e.g. LIWC, Pennebaker et al., 2001), but merely serve as the fix point of the concept in semantic space, to which each document is geometrically compared. Thus, rather than be all-encompassing, a dictionary needs to be as specific as possible, i.e., most closely match the "core" of the concept it is intended to represent. DDR has been demonstrated to outperform traditional word count methods in complex tasks such as classifying sentiment of movie reviews (Garten et al., 2018), or analyzing the moral loading of tweets (Hoover et al., 2018). Therefore, this method was ideally suited for analyzing the YouTube comments that social media users had produced in response to Logan Paul's "suicide forest" scandal, which we describe next. We provide a more detailed description about how DDR works, and advanced established techniques such as LIWC in the Supplementary Materials.

Method

Data Acquisition and Preliminary Cleaning

We scraped all comments on eight Logan Paul videos using the YouTube API .[1] Seven videos had been uploaded in the week directly prior to the scandal (one video per day between 22nd to 28th December 2017; comments were scraped on 27th January, 2018). These pre-scandal videos established the baseline attitude and commitment toward Logan Paul.[2] The post-scandal video was his apology (uploaded on 2nd January 2018; comments were scraped on January 21st, 2018).

First, top spam accounts were removed. Although YouTube has sophisticated algorithms for detecting and deleting spam (e.g., extensive self-promotion and advertisements, or links to third party sites), they do not identify all of it (Benevenuto et al., 2009). We therefore inspected the top contributing authors, (i.e., who had written more than 50 comments), and determined whether content was exclusively self-promoting or other spam, defined as comments entirely irrelevant to the video content. This resulted in the exclusion of 12 authors.

At this point, the dataset comprised 1,046,946 comments by 793,158 authors. Of those, 223,201 comments from 169,613 authors were from the seven pre-scandal videos, and 823,745 comments by 666,636 authors were from the post-scandal apology video. Because the main analysis required authors to comment both before and after the scandal, we matched the pre-scandal commenters with the post-scandal commenters based on their unique YouTube ID. This dataset contained 45,760 unique authors who had written 135,851 comments.

Pre-Scandal and Post-Scandal Measures

To assess the pre-scandal attitudes toward Logan Paul, three measures were used. As a behavioral measure of engagement with his content, pre-scandal *Engagement* was operationalized as the number of videos a user had commented upon during the week before the scandal. Because there were seven such videos, this count measure ranged from 1–7. We used frequencies as opposed to the number of comments per author because consistently commenting on each video can be considered a better indicator of in-depth engagement than occasional bursts. Indeed, the period included Christmas Eve, Christmas Day and Boxing Day, so taking time to watch and respond to YouTube videos over this holiday period reflected that an individual was a committed follower.

1) Replies to other people's comments were not considered because they depend on the original comment while we were interested in followers' direct response to the videos, rather than their response to other users.

2) Comments posted on those videos on or after 31st Dec, the day the "suicide forest" video was uploaded, were excluded.

To measure people's general *Liking* toward Logan Paul, we conducted a sentiment analysis for comments on the pre-scandal YouTube videos. We opted for this relative general valence measure because content varied widely, making the comments highly context-dependent. As a third pre-scandal measure, we operationalized *Social Identity* as the degree to which an author made use of Logan Paul specific fan language (e.g., "Logang4life"). Similar to Liking, the use of Logan Paul brand words should be relatively context independent.

To examine responses to the Logan Paul scandal (i.e., post-scandal attitudes), we categorized comments on his apology video regarding whether they expressed general *Support*, and the specific emotions of *Anger*, *Disgust*, *Adoration* and *Compassion* toward Paul.[3] We also again examined *Social Identity* as we did for pre-scandal attitudes.

DDR Preparation and Validation

Training and validation of the DDR classifier followed an intricate protocol. Here we give a short overview over the steps involved, a more detailed description is given in the Supplementary Materials. First, we hand-coded a randomly drawn subset of 1,000 comments for all dimensions, Average Cohen's Kappa (unweighted) = 0.79, Percent Agreement = 92%. Next, we defined concept dictionaries for all dimensions by identifying representative words from the hand-coded comments for each dimension, and used a word vector model generated from all collected YouTube comments to generate vectors for each dictionary. Finally, we used logistic regression with cross validation to train a classifier predicting the hand-coded comments from the dictionary vectors. We used F1 scores as the main measure to evaluate the performance of our models because they are commonly used for evaluations of binary classifiers (e.g., supporting or not supporting Logan Paul), and are calculated as the harmonic mean of precision and recall (Powers, 2011).

Performance was very good, with F1 scores ranging from 0.80 (Compassion) to 0.93 (Disgust). In addition, other common measures of classifier accuracy such as sensitivity, specificity, and Receiver Operating Characteristic (ROC) values were equally positive (see Table 1). Most importantly, the Support dimension could be reliably identified (F1 = .83).

3) The post-measures were somewhat different from the pre-measures because not all categories were equally meaningful at both time points. In particular, specific feelings of anger, disgust, etc. only became relevant after the scandal, whereas a broader positive-negative sentiment was more appropriate before the scandal. We also coded some more exploratory dimensions, namely whether comments expressed Forgiveness, Demand for further apologies, Relief, Justification attempts, or explicit Moral Character Judgments. These latter dimensions were included on a more ad hoc basis, and frequencies were overall low. Although we did not conduct a thorough analysis of these dimensions, the coded data is included in the Supplementary Materials.

Table 1

Model Performance for Variables of Interest

Variable		F1	ROC	Sensitivity	Specificity
1	Support	0.83	0.91	0.87	0.82
2	Adoration	0.85	0.88	0.81	0.82
3	Compassion	0.80	0.84	0.68	0.82
4	Social Identity	0.89	0.93	0.82	0.85
5	Anger	0.86	0.89	0.80	0.83
6	Disgust	0.93	0.96	0.87	0.91
7	Unclassifiable	0.91	0.85	0.87	0.66

Note. Model performance scores for each outcome variable, indicating performance of logistic regression models classifying comments on the post-scandal video. Scores closer to 1 reflect better performance.

A few final steps were taken before the main analysis. First, we used DDR to identify Unclassifiable comments in the post-scandal dataset. This led to the exclusion of 9,307 authors whose responses could not be categorized along the relevant dimensions, leaving 36,464 unique comment authors. Unless otherwise noted, all analyses were performed on this dataset.

Subsequently, we prepared the measures for analysis. Because the unit of analysis was the comment author and not the individual comment (i.e., we were interested in whether a given person supported Paul after the scandal), and the dataset was large and unbalanced (with different numbers of repeated measures for both independent as well as dependent variables, resulting from comment authors writing multiple comments both before and after the scandal), we averaged all dictionary loadings for each dimension for each author. Thus, each author had an individual Engagement score, pre-scandal DDR loadings for Liking, and Social Identity, as well as individual post-scandal loadings for Support, Anger, Disgust, Adoration, Compassion and Social Identity. Finally, the post-scandal loadings were converted into binary variables (i.e., Support vs. No Support, presence vs. absence of Anger, Disgust, etc.) using the pre-trained logistic regression models from the validation test.[4]

4) DDR loadings do not necessarily indicate the intensity/strength of a given dimension. Because we did not code for intensity in the human coded gold-standard data set, but treated each dimension as binary (i.e. present – absent) we cannot assess whether higher loadings indeed translate to higher intensity of the domain in question. Thus, our analysis used the validated classification method. We do however provide robustness analyses in the Supplementary Materials, which fall in line with the main models.

Results

Descriptive Statistics

We first examined the general composition of the dataset, starting with an overview of the *pre-scandal measures*. Overall, Engagement scores were positively skewed: of the 36,464 comment authors, ~73% (26,569) had an Engagement score of 1 (i.e., they commented on only 1 of the 7 pre-scandal videos), ~18% (6,509) an Engagement score of 2, and only ~9% (3,386) had an Engagement score of 3 or above. This indicates that consistent commenting was relatively rare.

To further understand the relationships between the pre-scandal measures of Engagement, Liking, and Social Identity, we computed their correlations. Because neither Engagement nor Social Identity DDR loadings were normally distributed (Engagement was right skewed, whereas Social Identity was left skewed), we used Kendall's tau. The Engagement score was significantly and negatively correlated with Social Identity (Kendall's tau = -.04, z = -9.18, p < .001), and Liking (Kendall's tau = -.01, z = -3.28, p = .0010). Thus, the number of videos users commented upon cannot be seen as a simple measure of positive attitudes towards Paul (although a visual inspection indicates a positive nonlinear relationship between higher values of Engagement and Liking, see Figure S1). Liking and Social Identity were positively correlated (Kendall's tau = .36, z = 103.69, p < .001), indicating that comment authors who wrote more positive pre-scandal comments also used more language relating to their fan identity.

Moving to the general distribution of the *post-scandal measures*, we again calculated overall percentages in the sample. Overall, 77% (27,990) of comment authors indicated continued Support for Logan Paul after the scandal, 60% (22,025) expressed Adoration in their comments, 44% (16,327) expressed Compassion, and 33% (12,168) expressed their Social Identity of being a fan, whereas only 16% (5,882) expressed Anger, and 4% (1,387) expressed Disgust. This indicates that the majority of Logan Paul viewers still saw Paul favorably after he apologized for his behavior.

To measure the co-occurrences between the post-scandal measures, we computed the phi-coefficient. This also serves as a validation check, identifying whether the DDR dictionaries were able to differentiate between the various dimensions. Indeed, Support for Logan Paul was negatively correlated with Anger, and Disgust, but positively correlated with Adoration, Compassion, and Social Identity (see Table 2). The strongest associations emerged between Support and Adoration, and Social Identity and Compassion.

Table 2

Correlations Between Post-Scandal DDR Scores

Variable	%	1	2	3	4	5
1. Support	77					
2. Adoration	60	.57***				
3. Compassion	44	.25***	.40***			
4. Social Identity	33	.34***	.54***	.43***		
5. Anger	16	-.33***	-.10***	.12***	-.06***	
6. Disgust	4	-.27***	-.13***	.02***	-.10***	.39***

Note. Correlations between the dimensions classified by DDR in post-scandal comments (N = 36,464). The Percentage column indicates the overall percentage of comment authors expressing the given dimension.
***p < 0.001.

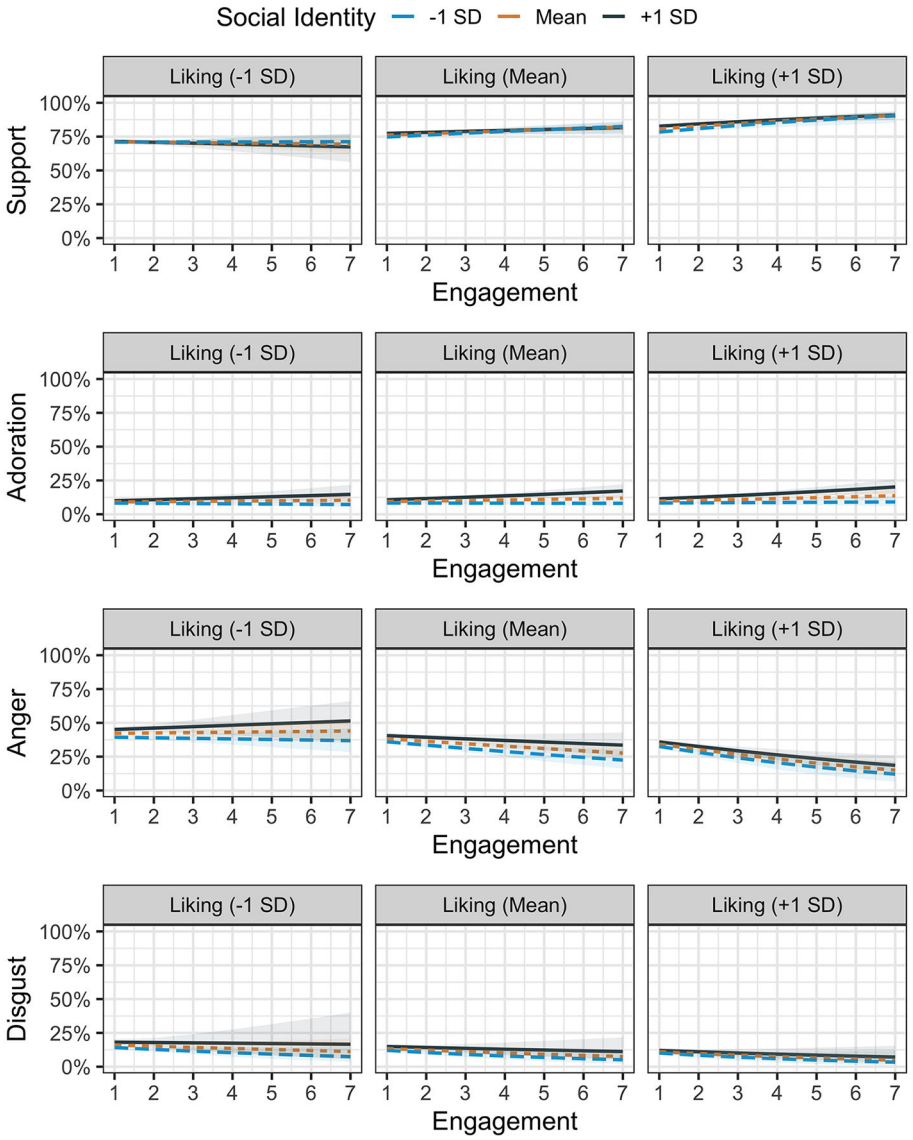
Testing the Link Between Pre-Scandal and Post-Scandal Attitudes

Support

Moving on to the key hypothesis, namely that YouTube users' attitudes toward Logan Paul would predict their support for him after the "suicide forest" scandal, we conducted a logistic regression, modelling post-scandal Support (supportive vs. non-supportive) with pre-scandal Engagement, Liking and Social Identity as independent variables, allowing for two-way interactions between the independent variables. Engagement significantly predicted post-scandal Support, suggesting that authors who had written more pre-scandal comments were more likely to still support him after Paul apologized for the scandal, Odds Ratio (OR) = 1.06, p < .001. Similarly, as hypothesized, Liking (OR = 3.52, p < .001) and Social Identity (OR = 1.10, p < .001) were positively related to Support. In addition, the interaction between Engagement and Liking was positive and significant, such that those users of frequently and positively commented on Paul's videos were more likely to show support (OR = 1.71, p < .001). In addition, Liking and Social Identity also significantly interacted, with users who both wrote positive comments and used social identifier terms being more likely to support Paul (OR = 1.54, p < .001). Thus, as hypothesized, all pre-scandal measures positively predicted post-scandal Support, with pre-scandal Liking being the strongest predictor (see Figure 1, and Table 3).

Figure 1

Predicted Probabilities for Support, Adoration, Anger and Disgust, Based on Social Identity, Liking, and Engagement, and Their Interactions



Note. Post-scandal Support, Adoration, Anger, and Disgust for Logan Paul, as predicted by pre-scandal Engagement, Liking, and Social Identity, and their interactions. Lines indicate predicted probabilities of the models in Table 3, ribbons depict confidence intervals. Plots for the other dimensions can be found in the Supplementary Materials (Figure S2).

Social Psychological Bulletin | 2569-653X
https://doi.org/10.32872/spb.8283

Table 3

Logistic Regression Results for Post-Scandal Variables

	Dependent Variable (post-scandal)					
	Support	Adoration	Compassion	Anger	Disgust	Social Identity
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	2.99***	0.10***	0.28***	0.67***	0.17***	0.04***
	[2.84, 3.14]	[0.09, 0.11]	[0.27, 0.30]	[0.63, 0.72]	[0.15, 0.19]	[0.04, 0.05]
Engagement	1.06***	1.04**	1.00	0.92***	0.90**	1.00
	[1.03, 1.09]	[1.01, 1.08]	[0.98, 1.03]	[0.89, 0.96]	[0.83, 0.97]	[0.97, 1.03]
Social Identity	1.10**	1.10**	1.13***	1.07	1.07	1.10***
	[1.03, 1.16]	[1.03, 1.17]	[1.07, 1.19]	[0.98, 1.15]	[0.90, 1.25]	[1.04, 1.16]
Liking	3.52***	1.12	0.68*	0.60	0.29*	0.93
	[1.92, 5.13]	[0.62, 1.63]	[0.43, 0.94]	[0.26, 0.94]	[-0.03, 0.61]	[0.56, 1.31]
Engagement * Social Identity	0.98	1.05*	1.07***	1.03	1.05	1.06**
	[0.94, 1.03]	[1.00, 1.09]	[1.03, 1.11]	[0.97, 1.09]	[0.93, 1.18]	[1.02, 1.10]
Engagement * Liking	1.71**	1.15	0.87	0.52**	0.77	1.23
	[1.15, 2.27]	[0.79, 1.51]	[0.64, 1.10]	[0.30, 0.73]	[0.13, 1.40]	[0.89, 1.57]
Social Identity * Liking	1.54***	1.25**	1.00	0.86	0.84	1.38***
	[1.34, 1.75]	[1.08, 1.43]	[0.88, 1.12]	[0.72, 1.00]	[0.59, 1.08]	[1.20, 1.56]
Support (post-scandal)		28.99***	3.73***	0.18***	0.06***	16.80***
		[26.74, 31.24]	[3.52, 3.94]	[0.17, 0.19]	[0.06, 0.07]	[14.94, 18.65]
Log Likelihood	-19,484.57	-18,069.24	-23,709.04	-14,369.71	-4,781.94	-20,380.95
Akaike Inf. Crit.	38,983.14	36,154.47	47,434.07	28,755.42	9,579.88	40,777.90

Note. Logistic regression results for post-scandal measures predicted from pre-scandal Engagement, Liking, and Social Identity, and their interactions. Estimates are odds ratios. Values in square brackets are 95% confidence intervals. $N = 36{,}464$.

*$p < .05$. **$p < .01$. ***$p < .001$.

Assessing the robustness of this finding, we built several alternative models: without specifying interactions of the independent variables, using the raw DDR scores instead of classified responses as the dependent variable, and adding additional predictors that could be extracted from the YouTube API. All model results fell in line with the main model (see Supplementary Materials, Tables S2–S5).

Subsequently, we conducted similar logistic regressions for all other post-scandal measures, again using pre-scandal Engagement, Liking and Social Identity, and their interactions, as predictors. Because post-scandal Support had been found to be correlated with most other dimensions, we controlled for it in these analyses. This allowed us to establish the predictive power of the pre-scandal measures over and above their ability to predict post-scandal Support. All model results can be found in Table 3.

Positive Emotions

Similar to the effect for Support, pre-scandal Engagement significantly predicted post-scandal Adoration, such that authors who had commented more often were more likely to express Adoration in their comments (OR = 1.04, p = .0052). Pre-scandal Social Identity showed a similar effect for Adoration (OR = 1.10, p < .001), and the interaction between Social Identity and Engagement was again positive and significant (OR = 1.05, p = .026). Pre-scandal Liking was not significantly related to adoration (OR = 1.12, p = .61), yet the interaction between Social Identity and Liking was again positive and significant (OR = 1.25, p = .016). Compassion, on the other hand, was not significantly related to pre-scandal Engagement (OR = 1.00, p = .86), but was negatively correlated with Liking (OR = 0.68, p = .047), and positively correlated with Social Identity (OR = 1.13, p < .001). In addition, the interaction between Social Identity and Engagement was significant and positive (OR = 1.07, p < .001).

Negative Emotions

Regarding the negative emotional responses investigated, pre-scandal Engagement predicted both Anger (OR = 0.92, p < .001) and Disgust (OR = 0.90, p = .062). For those two emotions, pre-scandal Liking was also a negative predictor, with authors who wrote more positive comments before the scandal being significantly less likely to express Disgust (OR = 0.29, p = .029). In addition, those users who were high in both Engagement and Liking also showed less signs of Anger (OR = 0.52, p = .0019).

Social Identity

Post-scandal Fan Language, as a reflection of social identity, was not significantly predicted by author Engagement (OR = 1.00, p = .99), mirroring the absence of such a correlation between these variables in the pre-scandal measures analysis. As expected, however, post-scandal Social Identity was positively associated with pre-scandal Social Identity (OR = 1.10, p < .001). This relationship was further qualified by a positive

interaction between Social Identity and Engagement (OR = 1.06, p = .0056), and a positive interaction between Social Identity and Liking (OR = 1.38, p < .001). This indicates that authors who identified themselves as part of the "Logang" prior to the scandal were also likely to still do so afterwards.

Overall, the general pattern of results is in line with the hypothesized positive relationship between pre-scandal Engagement, Liking, and Social Identity and post-scandal Emotions. Looking at effect sizes, the effect of Engagement was relatively small, whereas prior Liking emerged as the strongest predictor for Support, Anger and Disgust.

Comparing Followers vs. Non-Followers

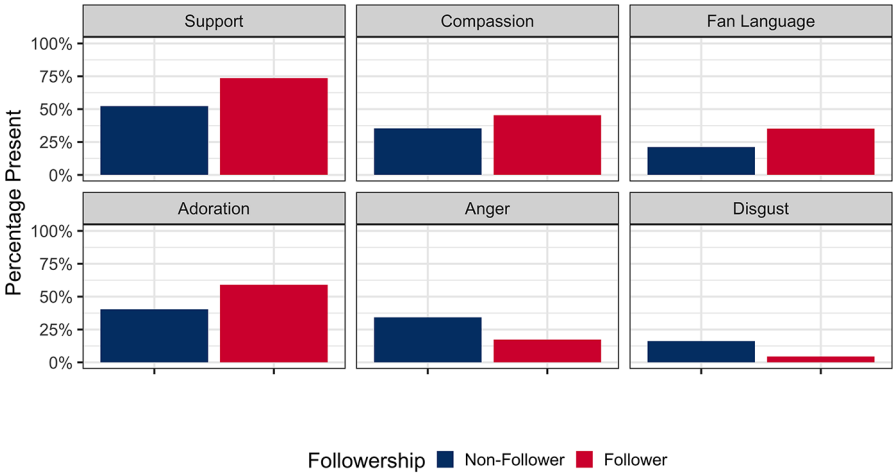
Lastly, we also compared post-scandal reactions of those authors who had commented at least once on a pre-scandal video to authors who only responded to the apology video after the scandal. The latter did not present a homogenous group, because they could either be people who had not heard of Logan Paul prior to the scandal, and where lured to this video by seeing discussions of it on other social media pages, the news, or people who were in fact followers but had not commented on any of the selected seven pre-scandal videos. Nevertheless, they formed an interesting comparison group, because they were clearly less highly engaged YouTube users (i.e., followers and non-followers), therefore providing additional insight into how pre-scandal attitudes predicted post-scandal reactions.

We categorized all comment authors present in the post-scandal video dataset based on their prior engagement, such that authors who had responded to at least one video were coded as followers, whereas authors who had not were coded as non-followers. This allowed an investigation of the complete dataset of the post-scandal video comments (N = 694,591).

Chi-square tests for each variable comparing the two groups revealed that authors who had shown engagement before the scandal showed more Support (74% vs. 52%), Adoration (59% vs. 40%), Fan Language (35% vs. 21%), and Compassion (45% vs. 35%) than authors who had not. In addition, they showed less Anger (17% vs. 34%), and Disgust (4% vs. 16%). All these differences were highly significant (p-values < .001). Thus, the two groups showed clearly distinctive responses to Logan Paul's apology video, with followers of Logan Paul being vastly more favorable towards the celebrity than non-followers (see Figure 2). When interpreting these results, it is important to keep in mind that because we only collected pre-scandal responses from seven videos, this likely underestimates the effect of followership, since some comment authors classified as non-followers might actually be regular Logan Paul viewers even if they had not publicly engaged with his content.

Figure 2

Percentage of Commenters Expressing Given Dimension by Followership



Note. Percentage of commenters expressing the given dimension (i.e., Support, Compassion, Fan Language, etc.), depending on whether they commented in at least one of the seven selected pre-scandal videos (Follower), or not (Non-Follower).

Discussion

This study investigated the role of prior positive attitudes about a celebrity on the degree to which followers forgave him once they were faced with a scandal involving a severe moral transgression. To do so, we took the Logan Paul "suicide forest" scandal as a case study, and measured expressed attitudes in both pre- and post-scandal (and post-apology) comments by using the DDR text analysis method (Garten et al., 2018). Linking these variables, we show that individuals who before the scandal had publicly exhibited displays of approval of Paul, shown greater general liking toward him, and identified most strongly as members of the Logan Paul fan community were also more likely to support him when he apologized after the scandal, and to largely (though note the exemption for Liking and Compassion) show positive (adoration and compassion) rather than negative (anger and disgust) emotional responses. This indicates that strong identification was fairly robust to the extremely negative light the scandal shone on Paul.

One of the most striking signs of this effect of prior positive attitudes on support arose from the comparison between followers and non-followers: Whereas followers of Paul had a generally positive stance towards the celebrity, the non-follower group was much more negative and seemingly unconvinced by his apology. Especially when considering that this analysis likely underestimates the effect of followership on post-scandal

attitudes, these results add weight to the idea that people can resist updating their beliefs about others when holding sufficiently strong prior beliefs.

This pronounced degree of support even in the face of extreme negative information speaks to the power of prior attitudes. Although many studies have shown a strong impact of negative information on the formation of attitudes (Cone & Ferguson, 2015; Rozin & Royzman, 2001), and the special role immorality has on the maintenance of social relationships (Fehr et al., 2010), we did not find a big influence of the scandal on Paul's followership. By contrast, our findings align more with research that has shown biased updating of beliefs in ways that fall in line with previously held beliefs. For instance, Hughes et al. (2017) showed that prior motivation influences the processing of information on the neural level, with participants failing to encode negative information about their ingroup. Similarly, research has provided evidence for biased processing of negative information regarding the attribution of blame or praise (Siegel et al., 2017), moral (Nadler & McDonnell, 2012), and even causal responsibility (Alicke et al., 2015). Therefore, our results are further testimony to the power of prior positive believes on attitude change.

Our findings are also in line with previous literature on social identity theory (Turner & Tajfel, 1986) showing that people who strongly identify with a group defend their social identity by denying group members' wrongdoings, and focusing their attention on positive aspects of the group (Branscombe et al., 1999). However, this theory and its findings have been largely developed and situated within political or religious groups, in which members know and regularly meet at least a part of their ingroup in person. In the case of Logan Paul, however, we can safely assume that the vast majority of his viewers have never met Paul, or interacted with him in any direct way. In fact, for many, this group may have solely existed online. Our study therefore extends previous findings to a new group, namely online fan groups, suggesting that they are guided by the same norms as more clearly established groups involving real-life contexts.

Online groups, and especially those forming on social media, are a growing part of today's social interactions, especially for young people. Recent findings on the phenomenon of echo chambers (Brady et al., 2017) showcase how groups may attempt to shield themselves from potential outside negative information. However, echo chambers have also been shown to sometimes be porous, allowing negative views to slip in (Turetsky & Riddle, 2018). Studying such online groups' behavior under identity threat has the potential to lead to new and more detailed insights into how groups and their members respond in these situations.

Importantly, our study adds to this literature by providing quantitative, real world, large-scale field data on a subject that is notoriously difficult to study (Finsterwalder et al., 2017). Previous studies investigating responses to scandals were often limited to mostly qualitative approaches, lacked pre-scandal data, or relied on experimental approaches using vignettes with uncertain external validity. Given ongoing debates

about the generalizability of psychological research (Yarkoni, 2022), providing such field evidence for laboratory-based findings is ever more important. Harnessing the power of machine learning and social media data, the current approach presents a solution for both problems. By demonstrating the usefulness of the DDR method (Garten et al., 2018) in approaching such research questions, we hope to have shown that it presents a versatile addition to the analysis of natural language, especially social media posts, that can extend traditional word count methods (e.g., Pennebaker et al., 2001). More specifically, DDR enabled us to develop customized concept dictionaries for post-scandal support that mapped more closely onto the dimension of interest than available word count dictionaries for sentiment or other basic emotions would have done.

On a broader scale, our results connect to recent Bayesian models of belief updating that provide a framework for understanding the mechanisms behind the seemingly irrational down-weighting of negative information. When a core belief is threatened and it is possible to create an ad-hoc hypothesis that can explain away the threatening information, no updating is necessary. Such an account explains, for example, why stereotypes can be so persistent: Individuals that do not fit the category can simply be further categorized into subgroups (Kim et al., 2020).

This also links up with the general idea of cognitive dissonance reduction (Festinger, 1962). As discussed above, witnessing an immoral act of a person that one holds dear can induce cognitive dissonance. Fans of Paul may therefore have been motivated to cope and reduce that dissonance, by either changing their attitude, or by finding ways to downplay the importance, or diagnosticity of the video on Paul's character (Uhlmann et al., 2015). Thus, positive moral evaluations of Logan Paul may have persisted in the face of counter-evidence if there was a good alternative explanation as to why the action was extrinsic to the actor's true self (Newman et al., 2015). For many followers it may have been less important whether Logan Paul's action was immoral or not, but what it said about his character. For example, one comment author wrote:

> "Logan, we've all made mistakes… From watching you for… Over a year now, I know your character, I know how you are, In that… You want to make content, you want to upload every second of your life to your viewers… The Logang, and, I also know your character, In that… You weren't trying to be malicious […]" (ellipses in original comment; comment id: UgxCU1IvZYVIEjIDYdV4AaABAg)

This follower expressed how their earlier character impression of Logan Paul was relevant to the interpretation of the new counter-attitudinal information. It thereby illustrates how a positive pre-scandal character judgment leads to evaluating the transgression as not intentional. This can be a very effective strategy, because non-intentional actions are not diagnostic when attempting to determine the true moral character of a person (Malle, 2011; Strohminger & Nichols, 2014).

Limitations and Constraints on Generalizability

Because it was a field study, a few limitations inevitably arose. First and foremost, this study cannot disambiguate the role of Paul's apology from direct responses to the scandal. Therefore, all shown effects are necessarily the effect of all of Paul's actions after the scandal.

Additionally, the case's and sample's inherent idiosyncrasies are important when interpreting the results. Logan Paul is a white man, conforming to a wide set of majority attributes. It is therefore possible that the amount of forgiveness Paul experienced would be diminished for a celebrity who has a minority background. As we do not have demographics information on commentators in our dataset, it is for future studies to test how such group identities interact with forgiveness after scandal.

In addition to that, as highlighted when discussing the role of disgust, Paul's wrongdoing falls into the specific domain of purity violations, which has been shown to be of particular importance for character evaluations (Giner-Sorolla & Chapman, 2017; Piazza et al., 2013). The here shown effects may therefore be limited to similar violations.

We were only able to investigate responses by individuals who decided to discuss certain content on a social media platform but had no information on those who refrained from doing so. Because the data analysis was based on matching authors, it therefore was impossible to gauge the reaction of those who did not choose to comment on Logan Paul's apology video. Thus, it could be that the observed pattern is due to more committed fans being more likely to silently withdraw from Logan Paul, without expressing this in the comment section. Although this is a possibility we cannot rule out, we deem this to be unlikely, because this group was the most vocal demographic pre-scandal, and many authors did not hold back expressing quite severe criticism. Moreover, given the anonymous context of the YouTube comment section, authors run little risk (e.g., social exclusion) for expressing views that would not resonate with the wider community.

Additionally, most of Paul's followers are young adolescents. Although specific data on his YouTube fan base is not available, it is fair to expect most of them to be members of the so-called Generation Z, who were born after the year 2000 (Turner, 2015). Thus, our results may be unique to this age group, because adults are usually more reflective, and thus may be more critical (Heckhausen et al., 1989). On the other hand, forgiveness has been shown to increase with age (Fehr et al., 2010), suggesting that our findings on young adults might in fact underestimate the actual magnitude of the effect. Moreover, similar effects of commitment and dissonance reduction have been shown in adults (Festinger, 1962; Forbes & Stellar, 2021). As a recent example, U.S. Americans who had voted for president Donald Trump did not reduce their support for him once they realized his tax reform had personally disadvantaged them; instead, they said that lower taxes were not all that important to them anyway (Rathje et al., 2020). Given this other literature, it is therefore unlikely that the effects observed here are only present in the age group we studied.

Another set of limitations arises from its embeddedness in social media. Because this study investigated reactions on social media, it is unclear how the observed effects translate to offline behavior. This point is commonly raised, and to date poorly understood (Crockett, 2017). Indeed, there are good reasons to expect that online behavior is special, because individuals often act anonymously, and may thus adopt different behavioral patterns or even personalities (Crockett, 2017). However, if online anonymity circumvents the social norms and constraints that people often find themselves in, one can speculate whether online behavior may in fact *more* accurately reflect a person's real opinions. Indeed, given the anonymous nature of the interaction, users might actually be more comfortable expressing negative views, therefore counteracting the social desirability bias that often comes with feeling evaluated by others.

Relatedly, one might argue that the case investigated here was special because the vast majority of Paul's followers did not actually know the celebrity in person. Thus, the observed effects may be limited to so-called parasocial interactions (Cerulo & Ruane, 2014). However, such relationships may actually not be all too different from normal relationships, particularly because social media allows for much stronger ties and more direct interactions with celebrities than previously possible (Cerulo & Ruane, 2014). This is especially the case for Logan Paul, who at the time of the scandal uploaded daily videos chronicling events from his life, frequently addressing his fanbase, arranging meet and greets, and other community events.

Future Directions

Given the demonstrated power of prior attitudes on forgiveness, one of the most important open questions concerns the issue as to when even highly identified fans will actually change their minds. In other words, although this study found a general positive effect for commitment on support, it is nevertheless still intuitively plausible that – under certain circumstances – commitment could lead to a radical change of heart and result in strong disapproval. Thus, there may be cases in which dissonance reduction fails, and the high attitude importance and affection will transform into strongly negative feelings and condemnation. Understanding these tipping points of character evaluations for highly committed individuals therefore represents an important path for future investigation.

This also relates to an uninvestigated mechanism in this study: the role of shame and guilt. Both emotions can be evoked by ingroup wrongdoings, as so-called forms of collective guilt or shame (Brown et al., 2008). Thus, an exploration of such emotions as indicators of updating, in addition to anger and disgust, could prove fruitful for the investigation of when and why people reverse even extreme positive attitudes.

Studying these kinds of relationships may require developing a feasible experimental paradigm in the laboratory. Personal relationships matter when it comes to moral judgments; Factors such as attachment and commitment influence the way in which people update attitudes (Kim et al., 2020), and make moral judgments (Nadler & McDonnell,

2012; Uhlmann et al., 2015). However, instilling a genuine sense of commitment in experiments may prove to be difficult. More research is therefore needed to understand the precursors of commitment, and how to study it in a controlled environment (Forbes & Stellar, 2021).

Finally, this study only looked at "original" comments, and disregarded any replies to them. However, it may be that the emerging discussions within the YouTube comment section reveal interesting patterns. For instance, Brady et al. (2017) found that moralized tweets spread more quickly and receive more attention on Twitter, but rarely escape their echo-chambers. Because the present dataset contains highly moralized content, and was drawn from a site where Logan Paul supporters and detractors come together, studying the inherent dynamics of the YouTube comment section might present new insights into how these two groups interact, and how moral outrage dynamically shapes online discussions.

Conclusion

When do we revise a previously positive character evaluation of another person given new negative information? This study shows that motivational factors make it an uphill battle. The more important an evaluation of a person is to us, the less likely this evaluation is to change given new unfavorable, counter-attitudinal information. As for Logan Paul, the "suicide forest" scandal was the start of a series of incidents that caused outrage in the wider social media sphere. Despite, or maybe even because of this, Paul's brand and his followership saw continued growth throughout. As our study indicates, Paul's success in surviving the scandal may in large part be due to his highly committed fanbase, the "Logang". For good reason, it seems, their catchphrase is "Logang for life".

Funding: Simon Karg was supported by the Santander Scholarship for the University of Cambridge.

Acknowledgments: We would like to thank Morteza Dehghani for providing highly valuable feedback on the implementation of the DDR method, and on an earlier version of the paper.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: For this article, data is freely available (Karg et al., 2020).

Supplementary Materials

The supplementary materials provide additional information, plots, and analyses for the main manuscript. In addition, we provide all code and data necessary to computationally replicate the analysis provided in the manuscript (for access see Index of Supplementary Materials below).

Index of Supplementary Materials

Karg, S. T., Lim, M., & Schnall, S. (2020). *Supplementary materials to "Followers forever: Prior commitment predicts post-scandal support of a social media celebrity"* [Data, Codebook]. OSF. https://osf.io/6xne8/

Karg, S. T., Lim, M., & Schnall, S. (2022). *Supplementary materials to "Followers forever: Prior commitment predicts post-scandal support of a social media celebrity"* [Additional Materials]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.7763

References

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives on Psychological Science, 10*(6), 790–812. https://doi.org/10.1177/1745691615601888

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., & Goncalves, M. (2009). Detecting spammers and content promoters in online video social networks. In M. Sanderson, C. Zhai, J. Zobel, J. Allan, & J. A. Aslam (Eds.), *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 620–627). The Association for Computing Machinery.

Bocian, K., Baryla, W., Kulesza, W. M., Schnall, S., & Wojciszke, B. (2018). The mere liking effect: Attitudinal influences on attributions of moral character. *Journal of Experimental Social Psychology, 79*, 9–20. https://doi.org/10.1016/j.jesp.2018.06.007

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America, 114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In R. Spears, N. Ellemers, & B. Doosje (Eds.), *Social identity: Context, commitment, content* (pp. 35–58). Wiley.

Brown, R., Gonzalez, R., Zagefka, H., Manzi, J., & Cehajic, S. (2008). Nuestra culpa: Collective guilt and shame as predictors of reparation for historical wrongdoing. *Journal of Personality and Social Psychology, 94*(1), 75–90. https://doi.org/10.1037/0022-3514.94.1.75

Cadinu, M. R., & Cerchioni, M. (2001). Compensatory biases after ingroup threat: 'Yeah, but we have a good personality'. *European Journal of Social Psychology, 31*(4), 353–367. https://doi.org/10.1002/ejsp.46

Canton, R. (2015). Crime, punishment and the moral emotions: Righteous minds and their attitudes towards punishment. *Punishment & Society, 17*(1), 54–72. https://doi.org/10.1177/1462474514548806

Cerulo, K. A., & Ruane, J. M. (2014). Apologies of the rich and famous: Cultural, cognitive, and social explanations of why we care and why we forgive. *Social Psychology Quarterly, 77*(2), 123–149. https://doi.org/10.1177/0190272514530412

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour, 1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

Fehr, R., Gelfand, M. J., & Nag, M. (2010). The road to forgiveness: A meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin, 136*(5), 894–914. https://doi.org/10.1037/a0019993

Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Stanford University Press.

Finsterwalder, J., Yee, T., & Tombs, A. (2017). Would you forgive Kristen Stewart or Tiger Woods or maybe Lance Armstrong? Exploring consumers' forgiveness of celebrities' transgressions. *Journal of Marketing Management, 33*(13-14), 1204–1229. https://doi.org/10.1080/0267257X.2017.1382553

Forbes, R. C., & Stellar, J. E. (2021). *When the ones we love misbehave: Exploring moral processes within intimate bonds*. PsyArxiv. https://doi.org/https://doi.org/10.1037/pspa0000272

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods, 50*(1), 344–361. https://doi.org/10.3758/s13428-017-0875-9

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond purity: Moral Disgust toward bad character. *Psychological Science, 28*(1), 80–91. https://doi.org/https://doi.org/10.1177/0956797616673193

Giner-Sorolla, R., Kupfer, T., & Sabo, J. (2018). What makes moral disgust special? An integrative functional review. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 57, pp. 223–289). Academic Press. https://doi.org/10.1016/bs.aesp.2017.10.001

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*(2), 366–385. https://doi.org/10.1037/a0021847

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Heckhausen, J., Dixon, R. A., & Baltes, P. B. (1989). Gains and losses in development throughout adulthood as perceived by different adult age-groups. *Developmental Psychology, 25*(1), 109–121. https://doi.org/10.1037/0012-1649.25.1.109

Hoover, J., Johnson, K., Boghrati, R., Graham, J., & Dehghani, M. (2018). Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra. Psychology, 4*(1), Article 9. https://doi.org/10.1525/collabra.129

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive and Affective Neuroscience, 12*(1), 49–60. https://doi.org/10.1093/scan/nsw147

Kim, M., Park, B., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences, 24*(2), 101–111. https://doi.org/10.1016/j.tics.2019.12.001

Malle, B. F. (2011). Attribution theories: How people make sense of behavior. In D. Chadee (Ed.), *Theories in social psychology* (pp. 72–95). Wiley-Blackwell.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111–3119.

Nadler, J., & McDonnell, M. H. (2012). Moral character, motive, and the psychology of blame. *Cornell Law Review, 97*(2), 255–304. https://doi.org/10.1016/j.tics.2019.12.001

Narvaez, D., & Lapsley, D. K. (2009). Chapter 8 Moral identity, moral functioning, and the development of moral character. *Psychology of Learning and Motivation, 50*, 237–274. https://doi.org/10.1016/S0079-7421(08)00408-8

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin, 40*(2), 203–216. https://doi.org/10.1177/0146167213508791

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Sciences, 39*(1), 96–125. https://doi.org/10.1111/cogs.12134

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Lawrence Erlbaum Associates.

Piazza, J., Russell, P. S., & Sousa, P. (2013). Moral emotions and the envisaging of mitigating circumstances for wrongdoing. *Cognition and Emotion, 27*(4), 707–722. https://doi.org/10.1080/02699931.2012.736859

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Rathje, S., Shariff, A., & Schnall, S. (2020). *The loyalty of the Trump voter: Continued support despite disappointing tax returns*. PsyArxiv. https://doi.org/10.31234/osf.io/bptqj

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Russell, P. S., & Giner-Sorolla, R. (2011). Social justifications for moral emotions: When reasons for disgust are less elaborated than for anger. *Emotion, 11*(3), 637–646. https://doi.org/10.1037/a0022600

Russell, P. S., Piazza, J., & Giner-Sorolla, R. (2013). CAD revisited. *Social Psychological & Personality Science, 4*(1), 62–68. https://doi.org/10.1177/1948550612442913

Schimkowitz, M. (2018). *Logan Paul's suicide forest video*. Know Your Meme. Retrieved 25.07.2018 from https://knowyourmeme.com/memes/logan-pauls-suicide-forest-video

Schnall, S. (2017). Disgust as embodied loss aversion. *European Review of Social Psychology, 28*(1), 50–94. https://doi.org/10.1080/10463283.2016.1259844

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition, 167*, 201–211. https://doi.org/10.1016/j.cognition.2017.05.004

Soukup, C. (2006). Hitching a ride on a star: Celebrity, fandom, and identification on the world wide web. *The Southern Communication Journal, 71*(4), 319–337. https://doi.org/10.1080/10417940601000410

Stoker, L. (1993). Judging presidential character: The demise of Gary Hart. *Political Behavior, 15*(2), 193–223. https://doi.org/10.1007/BF00993853

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Thomson, A. L., & Siegel, J. T. (2017). Elevation: A review of scholarship on a moral and other-praising emotion. *The Journal of Positive Psychology, 12*(6), 628–638. https://doi.org/10.1080/17439760.2016.1269184

Turetsky, K. M., & Riddle, T. A. (2018). Porous chambers, echoes of valence and stereotypes. *Social Psychological & Personality Science, 9*(2), 163–175. https://doi.org/10.1177/1948550617733519

Turner, A. (2015). Generation Z: Technology and social interest. *Journal of Individual Psychology, 71*(2), 103–113. https://doi.org/10.1353/jip.2015.0021

Turner, J. C., & Tajfel, H. (1986). The social identity theory of intergroup behavior. *Psychology of Intergroup Relations, 5*, 7–24.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Uhlmann, E. L., Zhu, L., & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry. *European Journal of Social Psychology, 44*(1), 23–29. https://doi.org/10.1002/ejsp.1987

Wojciszke, B. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology, 16*(1), 155–188. https://doi.org/10.1080/10463280500229619

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences, 45*, Article e1. https://doi.org/10.1017/S0140525X20001685

PSPS Polish Social Psychological Society

Social Psychological Bulletin (SPB) is an official journal of the Polish Social Psychological Society (PSPS).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.