Changes in online moral discourse about public figures during #MeToo

Benjamin M. Silver*, Kevin N. Ochsner
Department of Psychology, Columbia University, New York, NY 10027, USA
*Corresponding author email: bms2202@columbia.edu
Silver ORCID: 0000-0002-4763-3281

Abstract: During the #MeToo movement, the perceived morality of public figures changed in light of sexual assault allegations against them. Here, we asked how these changes were influenced by the perceived severity of alleged actions and by how well-known and well-liked were the public figures. Perceived morality was assessed by measuring (im)moral language usage in 1.4 million Tweets about 50 male public figures accused of sexual assault. Using natural language processing to analyze the tweets, we found that liking of public figures mitigated perceived immorality for less severe allegations, but had little effect on perceived immorality for more severe allegations. The persistence of negative perceptions one year later was related to liking and familiarity for the public figure, not allegation severity. These results suggest that in real-world contexts, we can forgive less harmful actions for people we like, but may not be able to if their actions are more harmful; over time, however, liking for others predicts lasting negative impressions of their moral misdeeds.

Keywords: social media, NLP, morality, person perception, belief updating

Statements and Declarations

Competing interests and funding: The authors have no competing interests or financial interests to disclose. The authors did not receive any external financial support to conduct this research.

Data Availability Statement: The tweets included in this study, as well as the statistical models, are publicly accessible at https://osf.io/z6c45/. The code for downloading and cleaning tweets and all statistical analyses is publicly accessible at https://github.com/bensilver95/metoo-twitter.

Author Contributions: Both authors made significant contributions to the theoretical ideas, methodological considerations, and analytic procedures in conducting the research, and both authors made significant contributions to writing and preparing this manuscript.

Acknowledgements: The authors wish to thank Niall Bolger for his advice on data analysis.

Activist Tarana Burke started the #MeToo movement in 2006. It went viral in 2017 and 2018 when over 250 (predominantly male) public figures were accused of committing sexual assault and/or abuse (North et al., 2020; Tambe, 2018). Many of these figures were previously revered and respected; as such, #MeToo provides a unique opportunity to study how the general public changes their discussions about public figures who are embroiled in public controversies. Here we ask whether and to what extent perceived morality of male public figures accused during #MeToo was influenced by prior familiarity with, and general liking of, the public figure, as well as the severity of the alleged actions.

At present, it is unclear how these variables may interact to cause an initial – or a lasting – change to population-level perceptions, in large part because relevant prior work has largely consisted of laboratory studies of how individuals change their beliefs about specific others. Together, these studies have shown that changes in beliefs about others happen if we receive evidence that initial attitudes or beliefs were incorrect or incompatible with subsequent behaviors the person in question demonstrates (Bhanji & Beer, 2013; Cone et al., 2021; Kovács, 2020; Mende-Siedlecki, 2018; Mende-Siedlecki & Todorov, 2016; Park & Young, 2020; Siegel et al., 2018).

However, for public figures – like politicians and Hollywood executives – the traditional approach of measuring shifting attitudes towards single individuals may not be the most useful level of analysis. Indeed, for public figures, it may be more important to study the ebb and flow of population-level discussions because they can determine large-scale outcomes such as who gets elected and what movies get made. While political science has long relied on public opinion polling to index population-level beliefs (e.g. Berinsky, 2017; Heath et al., 2005), here we took cues from psychological research on motivation and person perception to understand changes in public discourse surrounding figures accused during the #MeToo movement. No psychological study to date has investigated the response to or moral discourse around #MeToo accusations. Specifically, we investigated public discourse surrounding male public figures only, as the #MeToo movement was largely seen as a reckoning for powerful men, in particular (Tambe, 2018), as reflected by the fact that < 2.5% of public figures accused during #MeToo were women (North et al., 2020).

To accomplish this goal, we leveraged Twitter as a key source of data (Kachen et al., 2021; Xiong et al., 2019). For many years, natural language approaches to analyzing word usage in written texts, such as Linguistic Inquiry and Word Count (LIWC), have proved useful for understanding psychological responses to events, including emotions, beliefs, and attitudes (Mohammad, 2016; Pennebaker, 1997). Recently, these methods have been used to draw inferences about what Tweets can tell us about emotional responses to natural disasters (Sisco et al., 2017), political events (Simchon et al., 2020), violent acts that become national tragedies (Doré et al., 2015), and COVID-19 (Abdo et al., 2021; Metzler et al., 2023).

There are, of course, limitations to using Twitter data, including inherent difficulties in determining who/what is the subject of a tweet and understanding how to interpret the spread of a tweet (Burton et al., 2021). That said, Twitter data can provide a unique window into attitudes and beliefs on a large scale and over long periods of time. It also allows us to move beyond laboratory studies that examine impressions for novel (and often fictional) individuals about whom participants have no prior beliefs or feelings or immoral actions that are hypothetical or relatively unharmful -– which, to date, has been the norm – and ask whether changes in impressions about real-world figures endure over time.

Drawing inspiration from prior research, we sought to test three hypotheses regarding what tweets may reveal about population-level changes in moral discourse during the #MeToo movement. First, just as moral beliefs change when encountering evidence of immorality (Mende-Siedlecki et al., 2013a; Park & Young, 2020), we hypothesized that immoral language in tweets about public figures would increase sharply after the public figure was accused of sexual assault, as compared to baseline. Second, we hypothesized that general liking of, and familiarity with, public figures would predict the magnitude of changes in immoral language use. Lab studies have shown that we are likely to forgive close others for immoral behaviors (McCullough, 2001), which would suggest that higher liking and familiarity would lead to smaller increases in immoral language. However, harmful actions from close others can also lead to feelings of betrayal (Couch et al., 2017), which would mean that higher liking and familiarity could lead to *larger* increases in immoral language use. The question for #MeToo figures was which of these two paths public discussions would follow, and further, whether moves toward apparent forgiveness or betrayal would depend on the severity of the sexual assault allegations. Third, we sought to determine whether observed changes in immoral language use would persist over both short and longer time scales. In line with work showing that changes in beliefs and moral outrage wane over time (Crockett, 2017; Ferguson et al., 2019), we expected that immoral language use would lessen both in the short-term (the three weeks immediately after initial allegations) and in the long-term (one year later). However, given the consequential and real-world nature of the events, we anticipated that immoral language use one year later would still be higher than at baseline.

Method

Public Figure Selection

A four-step procedure was used for generating a list of public figures that met specific selection criteria. First, we began with a comprehensive list of 262 individuals accused of sexual assault in the #MeToo movement as compiled by Vox.com (North, 2019). Second, within this set of 262, we focused on individuals for whom initial allegations became public during a one-year span beginning on October 5th, 2017, the day that Harvey Weinstein's allegations became public (Kantor & Twohey, 2017). That date is widely seen as launching the #MeToo era (Kachen, 2021). While the one-year cut-off is somewhat arbitrary, the vast-majority of high-profile #MeToo cases emerged during this one-year period (only 7 public figures in the Vox database have cases that emerged after the one-year cut-off), and focusing on sexual assault allegations during the #MeToo era lends a degree of consistency and shared context to the data. This cut-off led to the exclusion of 24 public figures. Third, all female figures (N = 6) were removed, as the #MeToo movement was perceived as being about powerful men, in particular (Tambe, 2018). This perception is borne out in that only 2.3% of public figures from the comprehensive Vox.com list were women. Fourth, we removed three public figures (Donald Trump, Brett Kavanaugh, and Roy Moore) whose allegations were tied to broader political events, as we predicted that discussion of these events would be present in tweets and would be confounded with the data relevant to our hypotheses. Finally, we removed one public figure (Nelly) whose name was commonly used in other contexts, which made it difficult to select tweets about sexual assault specifically. This left a list of 228 public figures.

For each of these 228 individuals, an initial set of candidate tweets was selected from the first day after allegations became public. All tweets were collected using the Python package Twint (Zacharias, 2018), which scrapes tweets using Twitter's search function. Only tweets that

included the full name of the public figure, or the name of the public figure without spaces, were collected to ensure that the tweet was about the public figure specifically and not the situation more broadly. Duplicate tweets were removed to reduce the influence of retweeted news articles. To ensure that we had enough tweets for each public figure to conduct robust analyses, public figures who were mentioned in fewer than 1,500 tweets on the first day were removed from the sample. Thresholding to improve the quality of Twitter data is a common practice (Murphy, 2017), although there is little agreement about what the exact threshold should be. For this study, the threshold of 1,500 tweets was chosen based on a number of factors. One was the bimodal distribution of tweet counts for the initial sample of public figures, where 1,500 tweets was a clear demarcation point. Above 1500 tweets, tweet number per public figure was distributed relatively evenly; by contrast, below 1,500 tweets, the majority of public figures had very low numbers of tweets. In addition, 1,500 tweets was a high-enough number to ensure that a) the included public figures were associated with significant and widespread discussion about their sexual assault allegations, and b) that there was enough text for each public figure's tweets to reliably analyze the data. This thresholding procedure resulted in a final list of 50 public figures *(Figure 1A)*.

Tweet Selection

For each of these 50 public figures, tweets mentioning that public figure's name (or their name without spaces) were collected across three time periods. To establish pre-allegation levels of moral language use in tweets about each public figure, *baseline* tweet collection was conducted for a 21-day period six months prior to the allegations. Using pre-allegation tweets as a baseline allowed us to control for the effects of the allegations and ensure that any change in immoral language was a result of the allegation and not a general feature of that public figure. To assess changes from baseline caused by allegations, *initial response* tweets were collected from the 21 days following the first public allegations about sexual assault. To investigate whether changes in perceived morality were maintained over longer time periods, tweets were collected from a 21-day period *one year after* initial public allegations *(Figure 1B)*.

Data processing

Perceptions of morality in tweets were calculated using the Harm-Vice sub-list of the Moral Foundations Dictionary (MFD) (Graham et al., 2009), a natural language processing dictionary containing words related to both moral and immoral situations and characteristics. Not all words in the Harm-Vice sub-list are directly relevant to sexual assault, but because it includes words like *harm*, *abuse*, and *cruel*, it is the component of the MFD that most directly relates to the harmful and immoral behavior central to the #MeToo movement *(Figure 1C)*. The most widely-used software package for computing word counts, the Linguistic Inquiry and Word Count program (Pennebaker et al., 2015), calculates a score by determining the percentage of words in a tweet that are found in a particular sub-list of words. As such, tweets were concatenated by day (within each public figure's set of tweets) before being run through MFD (Tumasjan et al., 2010).

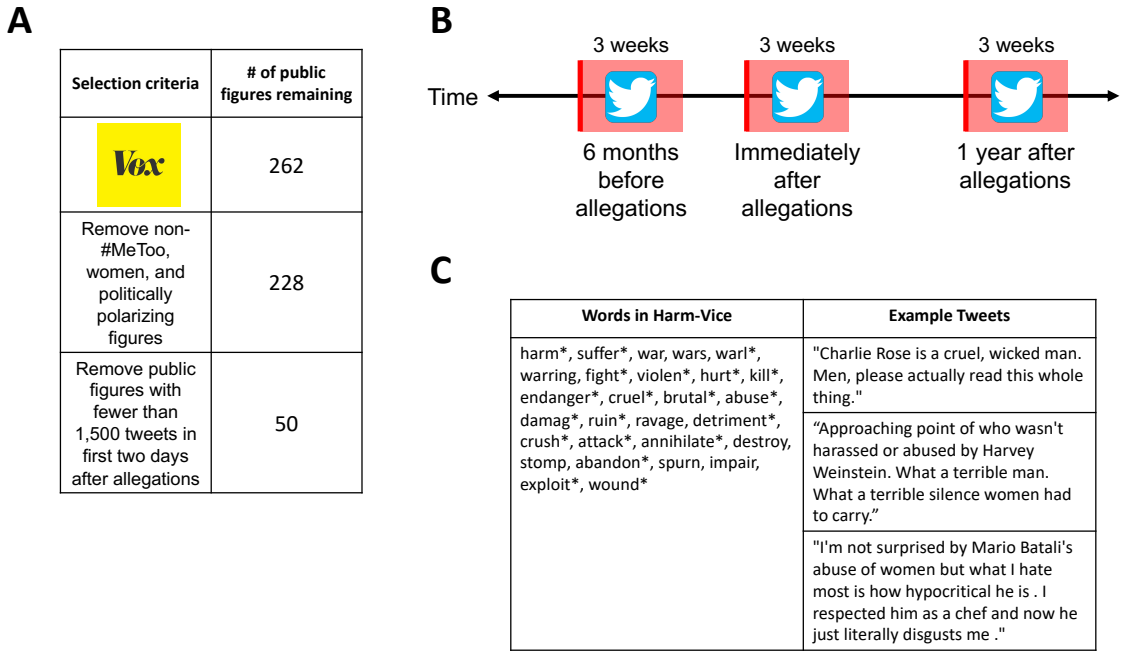

Fig. 1
Tweet selection methodology. Public figures were selected from a Vox.com database of public figures accused of sexual assault during #MeToo (A). We removed: public figures whose accusations emerged before October 5th 2017 or after October 4th 2018; the small number of female public figures included in this database (n = 6); male public figures tied to unrelated political events; and one male public figure whose name was difficult to search for in tweets. We then removed male public figures who were mentioned in fewer than 1,500 tweets the day after their allegations emerged. Tweets were collected using the Python package Twint for three weeks following allegations, for a three-week period six months before allegations, and for a three-week period one year after allegations (B). We used the Harm-Vice sub-list of the Moral Foundations Dictionary (C).

Data cleaning

For each collection period, three steps were taken to ensure that tweets accurately represented the conversation surrounding each public figure. First, all non-English tweets were removed. Second, all URLs were removed from the tweet text to reduce noise during linguistic analysis. Third, duplicate tweets were removed to reduce the influence of syndicated news articles often tweeted by bots. The resulting final tweet dataset consisted of 1,412,680 tweets, which included tweets from 6 months prior to allegations, tweets in the first three weeks after the allegations, and tweets from one year after the allegations. The median number of tweets per public figure was 10,281, with five public figures accounting for nearly half of the total number of tweets. See *Table 1* for summary data showing date of first public allegations and number of tweets at each time period for each public figure included in the study.

Table 1.
Public Figures and Tweets included in dataset

Public Figure	Date of Allegation	6 months prior to allegations	Initial 3 weeks after allegations	1 year after allegations
Al Franken	11/16/17	10841	137350	3400
Alex Jones	2/28/18	18736	26983	31191
Andy Dick	10/31/17	490	3361	429
Aziz Ansari	1/13/18	1470	28572	443
Ben Affleck	10/10/17	2703	19761	3836
Bob Weinstein	10/17/17	3	2072	43
Brett Ratner	11/1/17	161	14306	65
Bruce Weber	1/13/18	361	3091	387
Bryan Singer	12/4/17	347	8058	611
Charlie Rose	11/20/17	1593	44274	764
Chris Hardwick	6/14/18	316	17336	275
Cody Wilson	9/19/18	135	4841	90
Dustin Hoffman	11/1/17	738	6519	504
Ed Westwick	11/7/17	1655	8866	149
Eric Greitens	1/10/18	343	3441	99
Eric Schneiderman	5/7/18	489	15684	105
Garrison Keillor	11/29/17	470	12531	243
George HW Bush	10/25/17	1639	13837	1319
George Takei	11/10/17	3752	15295	2198
Glenn Thrush	11/20/17	767	4405	126
Harvey Weinstein	10/5/17	423	265282	7850
James Franco	1/11/18	6152	21013	1825
James Levine	12/3/17	97	3228	60
James Toback	10/22/17	5	7532	12
Jeremy Piven	10/31/17	246	3897	132
John Conyers	11/20/17	153	30419	149
John Lasseter	11/21/17	182	6381	134
Junot Diaz	5/4/18	276	4026	91
Kevin Spacey	10/29/17	1041	131284	4133
Les Moonves	7/27/18	226	12175	244
Louis CK	11/9/17	1980	73558	1259
Mario Batali	12/11/17	512	9472	158
Mark Halperin	10/25/17	289	10067	65
Marshall Faulk	12/11/17	375	2877	529
Matt Lauer	11/29/17	483	111214	1142
Morgan Freeman	5/24/18	5015	47761	393
Morgan Spurlock	12/14/17	185	3353	57
Oliver Stone	10/12/17	2014	4438	594
Roy Price	10/12/17	236	5041	606
Ryan Lizza	12/11/17	48	2119	38
Ryan Seacrest	2/26/18	1074	13786	1428
Scott Baio	1/27/18	2139	10729	953
Stan Lee	1/9/18	7272	12002	10008
Steve Wynn	1/27/18	244	24836	686
Steven Seagal	11/9/17	901	4162	758

Sylvester Stallone	11/16/17	1679	4638	2717
Tavis Smiley	12/13/17	145	7474	67
TJ Miller	12/19/17	2335	4161	751
Tom Brokaw	4/26/18	642	9181	218
Trent Franks	12/7/17	154	9083	42
Total		**83532**	**1245772**	**83376**

Operationalization of factors that may motivate changes in immoral language use

We used a combination of surveys and lexical analyses to provide estimates of three factors that may affect the way in which people tweet about the #MeToo allegations levied against public figures – liking and familiarity for each figure prior to allegations being made, and the severity/harmfulness of the alleged actions. As described below, each factor was measured in multiple ways in order to make our measurements more robust, and to include both subjective and objective methods of measurement.

Liking: Pre-allegation liking was calculated in two ways: a dictionary-based approach and a machine learning approach. For the dictionary-based approach, the sentiment analysis tool AFINN (Nielsen, 2011) was used. AFINN scores each word in a text as either negative (scores ranging from -3 to -1), neutral (0), or positive (scores ranging from +1 to +3). Examples of negative words include evil (-3), awkward (-2), and demanding (-1), while examples of positive words include lenient (1), inspirational (2), and great (3). Tweets were concatenated by day and public figure, and AFINN scores were normalized based on the length of the text concatenation. For the machine learning method, a binary classification transformer model, using the uncased DistilBERT model (Sanh et al., 2020), was then trained on a dataset of 160,000 tweets from the *Sentiment140* dataset, which were each classified as either positive (+1) or negative (0), through the Simple Transformers Python package (Rajapakse, 2020). This model was then run on all of the baseline tweets, with each tweet classified as either positive or negative. Liking of the public figure was measured by averaging the transformer model score across all baseline tweets for that public figure.

Familiarity: Familiarity with the public figure was measured in four ways. The pygooglenews python package (Burgara, 2020) was used to measure trending news headline mentions of each public figure in the same 21-day period before each allegation, and number of tweets that mentioned the public figure was measured over this same 21-day period as well. These two measures were strongly correlated with each other ($p < .001$), suggesting that pygooglenews is a valid index of general news trends online.

Finally, we recruited 80 online survey participants (age range: 18-65) via Prolific to assess the prominence and power of each accused public figure. Prominence refers purely to fame – how well is the public figure known by the general population? Power refers to level of influence, which can be bestowed by money or social status. Prominence and power were separated in this survey, since some important cases during the #MeToo movement concerned public figures who were not necessarily household names before their accusations emerged, but nonetheless held significant power over others. Detailed definitions of prominence and power were shown to participants at the start of the survey. During the survey, the participants were shown the name and a photo of each public figure, and asked to retrospectively estimate, on a 0-100 scale, the prominence and power of each public figure before their sexual assault allegations emerged. (Participants were also asked to estimate current, post-allegation power and prominence, but these ratings were not used for analyses.)

Severity: Severity of the allegation was measured in two ways. First, each allegation against a public figure was summarized and anonymized, and shown to 100 participants recruited online via Prolific (age range: 18-65), who rated it on a 0-100 scale in terms of the amount of harm the event/behavior caused. Second, we created a rubric with four dimensions: number of people affected, length of time of behavior, type of behavior, and context. We scored each allegation against each of these dimensions, allowing us to calculate a summary severity score.

Prior to data analysis, the individual components for each factor were scaled around zero and averaged, so that there was one score for each factor.

Determining heterogeneity

In order to justify analyses of factors that may have motivated changes in moral discourse for public figures, we sought to demonstrate that there was indeed significant heterogeneity in the individual differences of changes across public figures. The SD of the slopes from six months prior to the first three weeks between public figures was 0.493 (95% CI: [0.394, 0.611]). Using two of the criteria laid out in Bolger et al. (2019), we determined that the heterogeneity of the slopes between time periods was significant. First, the random effect's 95% confidence interval did not surround 0 ([0.394, 0.611]), suggesting that the effect is likely not due to sampling error. Second, it is recommended that the random effect be larger than 25% of the fixed effect, and we found that in the present model it was equal to roughly 72% of the fixed effect (RE = 0.493, FE = 0.684, | RE/FE | = 72.07%). Despite this heterogeneity, 47/50 of the within-public figure slopes were positive, meaning that immoral language use increased, while the remaining three had 95% CIs surrounding 0.

Results

Did #MeToo allegations lead to a change in immoral language use?

Our first hypothesis concerned the immediate effects of sexual assault allegations on immoral language use online. We ran a multi-level Bayesian model with random intercepts and slopes, with public figure as the nesting variable (essentially, each public figure was treated as a study participant) and time period (a baseline period 6 months prior to allegations vs the first three weeks following allegations) as a random effect. Our model revealed that 0.16% of words in each day of the baseline tweets were found in the Harm-Vice list from the MFD – hereafter referred to as immoral language – with a between-public figure SD of 0.115 (95% CI: [0.066, 0.165]). In addition, there was a fixed effect of tweet time period, for which immoral language on each day were higher in initial response tweets as compared to baseline tweets (b = 0.683, SE = 0.071, 95% CI = [0.539, 0.823]), meaning that, on average, immoral language post-allegation quintupled as compared to baseline *(Figure 2)*.

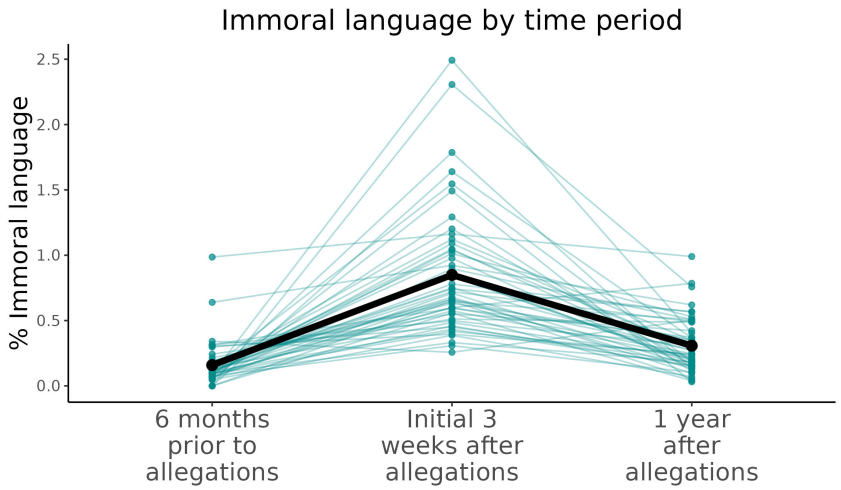

Fig. 2
Average immoral language use for tweets in a three-week baseline period six months before allegations, a three-week period immediately after allegations, and a three-week follow-up period one year after allegations. Immoral language in tweets was calculated using the Moral Foundations Dictionary Care-Vice sub-list of words. Each line represents a public figure. The thick black line is the mean.

What factors predicted changes in immoral language use?
To address this question, we ran a Bayesian, multi-level model, with three potential factors – allegation severity, liking, and familiarity – as interacting predictor variables, and controlled for levels of immoral language in the baseline tweets figure (*Figure 3A*). We found a significant positive effect of allegation severity (b = 0.279, SE = 0.105, 95% CI = [0.074, 0.483]), meaning that more severe actions led to tweets with more immoral language. We did not find a main effect of liking (b = -0.058, SE = 0.130 95% CI = [-0.320, 0.194]) or familiarity (b = -0.098, SE = 0.104, 95% CI = [-0.295,0.120]). However, there was an interaction between liking and allegation severity (b = 0.562, SE = 0.246, 95% CI = [0.085, 1.037]), such that at low severity levels, higher liking led to less immoral language use, while at high severity levels, higher liking did not predict a difference in immoral language use, with the interaction trending towards slightly *more* immoral language (*Figure 3B*). There were no significant interactions between any of the other factors. Together, these results demonstrate that the severity of the action was most important in predicting the overall amount of immoral language following the #MeToo allegation, but that this effect differed as a function of liking for the public figure: for well-liked figures, tweets about them saw little change in immoral language use if the alleged actions were perceived to be less severe; by contrast, immoral language use increased significantly for severe allegations.

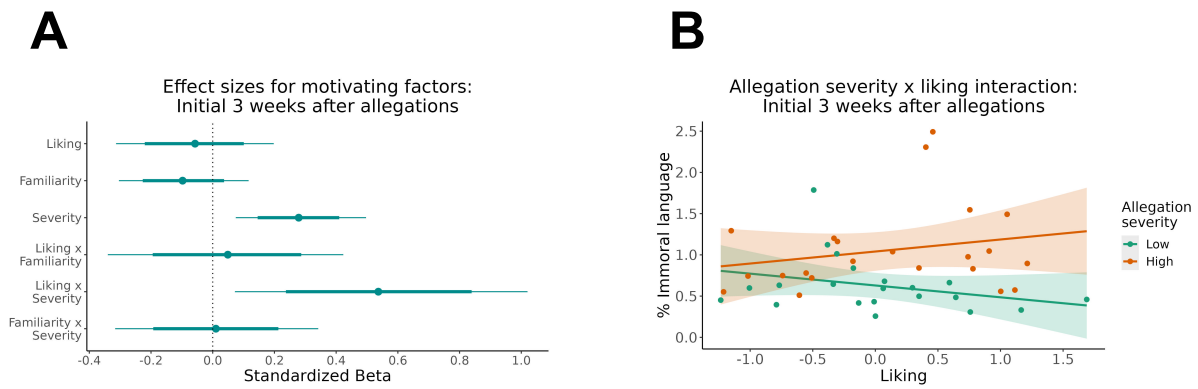

Fig. 3
Effects of liking, familiarity, and allegation severity for initial three weeks. The effect sizes for each motivating factor's effect on overall immoral language in the initial three weeks, as defined by the MFD Care-Vice sub-list of words, are shown in (A). Thick bars are 80% credibility intervals and thin bars are 95% credibility intervals. The interaction between liking and allegation severity in the initial three weeks is shown in (B). Each point represents the average amount of immoral language in tweets about a public figure. Ribbons are 95% CIs.

What were the temporal dynamics of changes in immoral language use?

Short-term effects: To address short-term effects, we conducted an analysis of how immoral language in tweets changed in the three weeks following the allegation, on a day-to-day basis. In this model, number of days after allegations (0-20) was a predictor variable, as both a fixed and random effect. We found an effect of day (b = -0.043, SE = 0.006 95% CI: [-0.055, -0.031]), in which the percentage of words classified as immoral decreased over time in the first three weeks, but not in the tweets from six months prior (*Figure 4*). The decrease in immoral language appears to be exponential rather than linear, so we re-ran the model with the logarithm of Harm-Vice scores, and found a similar effect (b = -0.038, SE = 0.006, 95% CI = [-0.049, -0.027]), suggesting that the majority of the immoral language drop-off occurs early on after sexual assault allegations occur.

Next, we removed time period as an interaction term and limited our data to the first three weeks. In a model with day and all three motivating factors as interacting fixed effects and day as a random effect, we failed to find any interactions between motivating factors and day (severity: b = -0.007, SE = 0.009, 95% CI = [-0.024, 0.011]; familiarity: b = 0.005, SE = 0.009, 95% CI = [-0.013, 0.023]; liking: b = -0.013, SE = 0.011, 95% CI = [-0.035, 0.007]). This suggests that while motivating factors affect overall levels of immoral language in the three weeks following an allegation, they do not modulate the trajectory of the decrease in immoral language over time. Given that the majority of the change occurs in the first week, we ran an identical model, with only data from the first week after each allegation. However, we still failed to find any significant interactions between day and motivating factors.

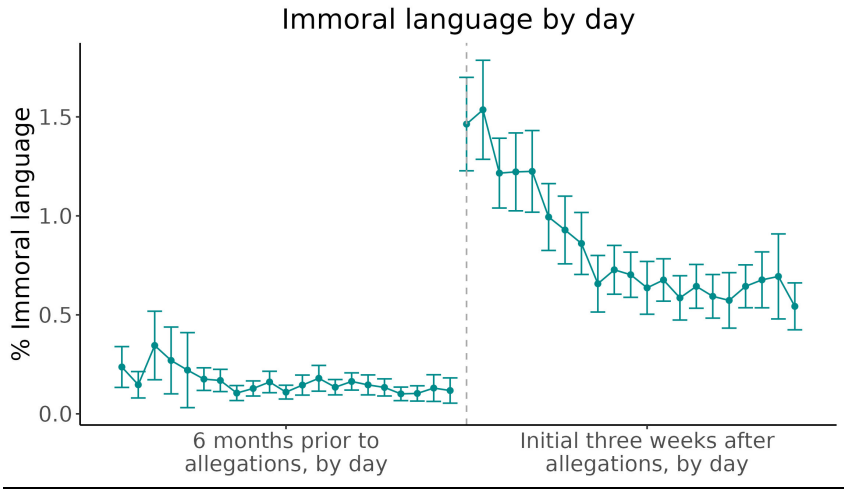

Fig. 4
A comparison of baseline vs updating period immoral language use by day. Each dot represents a score for one day, with scores relatively constant for baseline tweets and decreasing over time for updating period tweets.

Long-term effects: To address long-term effects, for all public figures, we collected all tweets that mentioned them in a three-week period exactly one year after their allegations. We found that even one year later, the amount of immoral language was significantly higher than at baseline (b = 0.141, SE = 0.023, 95% CI = [0.097, 0.186]), but was significantly lower than in the three weeks immediately following an allegation (b = -0.541, SE = 0.070, 95% CI = [-0.680, -0.406]) (*Figure 2*).

We next ran an identical Bayesian multi-level model on tweets from the one year later period. While allegation severity was more predictive of immoral language in the first three weeks, in the tweets from one year later, familiarity and liking were more predictive than severity was (liking: b = -0.067, SE = 0.060, 95% CI = [-0.189, 0.053]; familiarity: b = -0.071, SE = 0.049, 95% CI = [-0.170, 0.025]; severity: b = 0.025, SE = 0.044, 95% CI = [-0.061, 0.119]), with higher levels of familiarity and liking predicting lower levels of immoral language (*Figure 5A*). While the credibility intervals for effects of familiarity and liking were not completely outside zero, this result suggests that the primacy of the action in motivating changes in immoral language is a proximal effect, and that the effects of target-related motivations on moral discourse are more persistent over longer time periods. However, the interaction effect between liking and severity that was present in the initial three weeks was no longer present (*Figure 5B*).

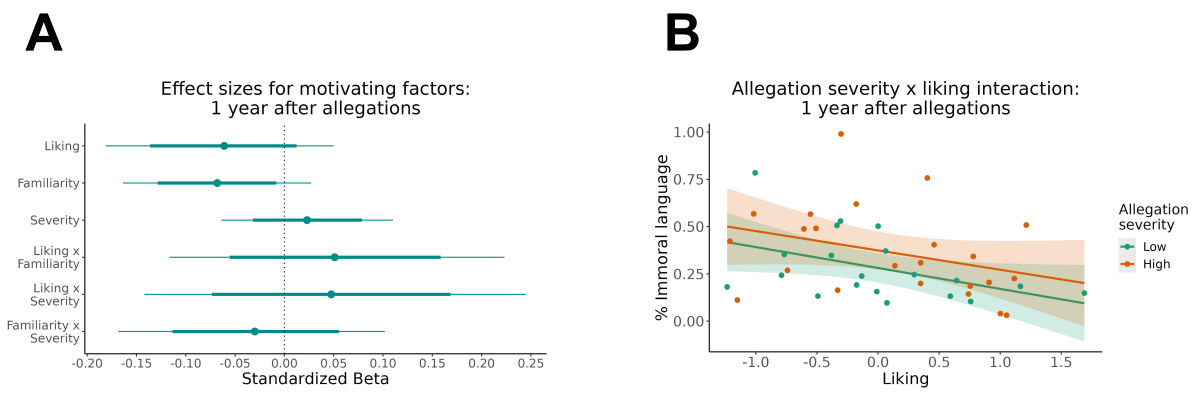

Fig. 5
Effects of liking, familiarity, and allegation severity for tweets one year later. The effect sizes for each motivating factor's effect on overall immoral language use one year later, as defined by Care-Vice, are shown in (A). Thick bars are 80% credibility intervals and thin bars are 95% credibility intervals. The interaction between liking and allegation severity at one year later are shown in (B). Each point represents the average amount of immoral language in tweets about a public figure. Ribbons are 95% CIs.

Discussion

The #MeToo movement is thought to have sparked significant changes in the public discourse about the morality of dozens of prominent male public figures (Tambe, 2018). Here, we provide the first empirical evidence about the nature and key predictors of these changes. At multiple time points we analyzed the moral language used in tweets about male public figures accused of sexual assault to investigate the factors that predicted population-level changes in moral discourse. Three key findings were observed.

First, we found that #MeToo allegations significantly increased immoral language use in tweets about accused figures, which fits with prior lab-based findings on impression updating (Mende-Siedlecki et al., 2013b; Reeder & Spores, 1983; Siegel et al., 2018) and responses to controversies on social media (Brady et al., 2021). Both types of research have shown that targets previously viewed as moral can be rapidly reassessed as immoral if we learn they have been accused of committing immoral acts. Among possible explanations for these findings, two are salient in the context of #MeToo: not only are immoral behaviors less common than moral ones, and therefore more diagnostic of a person's character (Fiske, 1980; Siegel et al., 2018), negative information generally is more impactful because it may signal a potential threat (Baumeister et al., 2001).

Second, the magnitude of the initial spike in immoral language depended on both the severity of alleged actions and how well-liked a public figure was before allegations emerged: liking mitigated immoral language for less severe allegations but trended towards an increase in immoral language for more severe allegations. This patterns suggests that we may collectively overlook, explain away, or forgive the immoral actions of liked individuals – so long as those actions don't seem too severe (Bradfield & Aquino, 1999; Fourie et al., 2020) – but for misdeeds of greater magnitude, then we may ignore our feelings of liking, or view them as increasingly immoral and even express moral outrage or feelings of betrayal (Couch et al., 2017; Couch & Olson, 2016). These findings also fit with lab studies showing similar effects when individuals evaluate social targets (Kihlstrom, 2013), including in the context of sexual harassment (Pryor et al., 1993).

Third, the level of immoral language in tweets one year after initial allegations was still greater than the pre-allegation baseline, but below the level seen during the three weeks immediately following the allegations. While this finding fits with studies showing that lasting changes in interpersonal beliefs happen only when an inconsistent behavior is both diagnostic and believable (Ferguson et al., 2019), it is important to note that whereas immoral language use was driven most strongly by allegation severity for the first three weeks, it was driven by liking for and familiarity with the public figures one year later. This pattern suggests that alleged actions provided an initial basis for moral discourse because they were highly accessible, concrete, and available to influence behavior (Doré et al., 2015; Higgins & Brendl, 1995; Rothbart et al., 1978; Schwarz et al., 1991). But after a while, the initial conversational focus faded, leaving the general public to base discourse on long-standing attitudes (such as liking and familiarity) towards a given figure. These findings fit with laboratory work showing that interpersonal beliefs can be change-resistant (Cao & Banaji, 2016), and often return to baseline quickly when changes occur (Lai et al., 2016). However, our findings go beyond this work by showing that over longer periods of time pre-existing attitudes toward people we know may be important and persistent predictors of collective judgments about them. In part, this may reflect the durability of semantic or "gist-like" representations of someone's traits, which we tend to rely on when making judgments about others (Klein et al., 1996; Sherman & Bessenoff, 1999; Wagner et al., 2019).

This leads to an important consideration – the public nature of the #MeToo movement means that there could also have been important situational and social influences on the emergence of sexual assault allegations. First, the phrase #MeToo existed for several years before the accusations against Harvey Weinstein; it is thought to have become a public and widespread movement because both the accused and the accusers had access to more public platforms (Tambe, 2018). It's precisely the public nature of this phase of the #MeToo movement that allowed us to analyze widespread conversation on social media. In this way, this method can also be considered a limitation, as it prevents us from analyzing how perceptions of moral character change for less prominent people accused of sexual assault. Second, the #MeToo movement may have involved a snowball effect: the more people who went public with allegations, the more comfortable others became with going public as well (Gallagher et al., 2019). Over time, it's possible this affected the general public's perceptions of survivors of sexual assault, which in turn could have influenced higher-level discussions each time new allegations emerged.

Similarly, the widespread coverage in conventional media channels and discussion on social media platforms could have influenced whether, when, and how a given person decided to tweet about one of the accused figures. The public nature of the #MeToo movement also may have influenced motivations to tweet, as it is known that on social media platforms people can be rewarded by their social networks for expressing moral outrage (Crockett, 2017), and that a relatively small number of users are responsible for a majority of posts (Brady et al., 2021). Many previous studies that use social media look at the spread of attitudes and information in an online context (Brady et al., 2017; Goldenberg & Gross, 2020; Schöne et al., 2021) by analyzing the salience of posts through reactions such as likes and retweets. In our study, we avoided these issues because we were not studying how information spreads, but rather, how population-level perceptions can be gleaned from aggregating social media data.

This naturally begs the question: Are these perceptions about the public figure, their alleged actions, or both? While counting immoral word usage in tweets can't differentiate these

possibilities, for two reasons we think it's possible that (im)moral language in tweets may more strongly reflect beliefs/attitudes about public figures. First, beliefs about the morality of negative actions are typically stable over time (Goodwin & Darley, 2012). As such, the post-allegation increase in tweet volume and immoral language might reflect increased discourse about the qualities (e.g., the moral character) of the person implicated. Second, prior studies suggest that any immoral action is inextricably tied to the actor's perceived morality (Gantman & Van Bavel, 2015), which suggests the tweets we analyzed may reflect population-level beliefs/attitudes about the morality of accused public figures. This logic has similarly informed prior Twitter studies about reactions to public events (Doré et al., 2015; Metzler et al., 2023; Schöne et al., 2021; Simchon et al., 2020) and even sexual assault accusations (Maryn & Dover, 2023).

In addition to some ambiguity regarding the subject of the immoral language, there were two other limitations to our dictionary-based method of analysis. First, dictionary-based methods are not able to detect sarcasm or irony, which are popular forms of expression on social media platforms like Twitter (Sykora et al., 2020). Although there has been some work seeking to detect sarcasm using machine learning (Sarsam et al., 2020), these approaches are inexact. While there may have been some sarcastic or ironic language present in our data, the sheer quantity of our data (over one million tweets) makes it unlikely that this language significantly influenced our results. Second, the dictionary that we used, the Harm-Vice sub-list from the Moral Foundations Dictionary, may not have fully encompassed all language that is relevant to discussions of sexual assault allegations. Choosing the correct list of terms is often a topic of debate in dictionary-based research, and one that can potentially allow for large researcher degrees of freedom. We exclusively used a pre-existing list, based on a well-researched psychological construct within Moral Foundations Theory (Piazza et al., 2019), to ensure that our findings would be replicable and based on existing psychological theories.

We should also note that there are several other factors we did not test that may have impacted the magnitude of observed changes in immoral language. First, as previously stated, we excluded female public figures from analyses. The conversation about sexual assault committed by women is of a distinct nature, with additional considerations regarding power and gender (Gannon et al., 2008). Given that there were only 6 women in this dataset, we did not have enough datapoints to meaningfully compare the discourse around female public figures to the discourse around male ones. A future study may wish to systematically test these differences. Beyond excluding female public figures from our analyses, the identity of the accused, as well as the identity of the accuser, was not included in our models. Future work on this issue could examine whether population-level perceptions and discourse may have been impacted by the race of the accused, as race can play a role in perceptions of moral character (Eberhardt et al., 2006; Stanley et al., 2011). We did not test factors related to race in the current study because 43 out of 50 of our public figures were white.

Public figures accused of sexual assault also came from a wide variety of professions, from politics to Hollywood. The general public likely has different baseline assumptions about the moral character of people from different professions, perhaps because of differential perceptions of power (For example, a Hollywood executive might be deemed to have more power than a journalist.) Preliminary analyses on a subset of our data revealed that immoral language in the first three weeks increased more for figures from Hollywood than for figures from journalism/media, although this effect may be confounded with allegation severity, which was higher for Hollywood figures than for any other profession. A future analysis may wish to systematically compare population-level discussions about sexual assault across professions.

Finally, our data at the one year later timepoint may have been impacted by the fallout from the allegations: some public figures may have released genuine, well-received apologies, while others may have denied the accusations, and still others may have been cleared of wrongdoing altogether. We did not systematically test "allegation outcome" alongside our measures of liking, familiarity, and allegation severity. However, the emergence of an accusation during #MeToo typically led to more widespread media coverage than ensuing apologies and legal proceedings; thus, we believe the impacts of allegation outcome on our results are minimal. Despite these limitations, it should be noted that our aggregate results still hold. Regardless of the race and occupation of the accused or the accuser, or the outcome of the allegations, the pattern of results found in the paper still emerges in aggregate. As such, our findings may represent an average effect across race and occupation. Future work may wish to see if the present effects hold, are exacerbated, or are mitigated for specific categories of accused public figure or accuser.

In sum, the present data remind us that even people we like and are familiar with may act in ways that challenge our preconceived notions about them. Do these moments pass by without impact or influence? Are they actively explained away? Or do they profoundly change our perceptions? The current study addressed this issue in the context of the #MeToo movement, asking how society reacts when public figures that we know and like are alleged to have committed immoral acts. Changes in tweet content suggested that changes in the moral discourse about public figures did indeed occur, and that the nature and persistence of these changes was dependent on both the severity of alleged actions as well as how well-liked and well-known was a given public figure. These results highlight that collective beliefs about public figures may be constantly in flux, influenced by our prior attitudes and beliefs as well as our perceptions of their actions.

Statements and Declarations

Competing interests and funding: The authors have no competing interests or financial interests to disclose. The authors did not receive any external financial support to conduct this research.

Data Availability Statement: The tweets included in this study, as well as the statistical models, are publicly accessible at https://osf.io/z6c45/. The code for downloading and cleaning tweets and all statistical analyses is publicly accessible at https://github.com/bensilver95/metoo-twitter.

Author Contributions: Both authors made significant contributions to the theoretical ideas, methodological considerations, and analytic procedures in conducting the research, and both authors made significant contributions to writing and preparing this manuscript.

Acknowledgements: The authors wish to thank Niall Bolger for his advice on data analysis.

References

Abdo, M. S., Alghonaim, A. S., & Essam, B. A. (2021). Public perception of COVID-19's global health crisis on Twitter until 14 weeks after the outbreak. *Digital Scholarship in the Humanities*, *36*(3), 509–524. https://doi.org/10.1093/llc/fqaa037

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Berinsky, A. J. (2017). Measuring Public Opinion with Surveys. *Annual Review of Political Science*, *20*(1), 309–329. https://doi.org/10.1146/annurev-polisci-101513-113724

Bhanji, J. P., & Beer, J. S. (2013). Dissociable Neural Modulation Underlying Lasting First Impressions, Changing Your Mind for the Better, and Changing It for the Worse. *Journal of Neuroscience*, *33*(22), 9337–9344. https://doi.org/10.1523/JNEUROSCI.5634-12.2013

Bradfield, M., & Aquino, K. (1999). The Effects of Blame Attributions and Offender Likableness on Forgiveness and Revenge in the Workplace. *JOURNAL OF MANAGEMENT*, *25*(5), 25.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

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*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Burgara, A. (2020). *Pygooglenews* (0.1.2) [Python]. https://github.com/kotartemiy/pygooglenews

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635. https://doi.org/10.1038/s41562-021-01133-5

Cao, J., & Banaji, M. R. (2016). The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences*, *113*(27), 7475–7480. https://doi.org/10.1073/pnas.1524268113

Cone, J., Flaharty, K., & Ferguson, M. J. (2021). The Long-Term Effects of New Evidence on Implicit Impressions of Other People. *Psychological Science*, *32*(2), 173–188. https://doi.org/10.1177/0956797620963559

Couch, L. L., Baughman, K. R., & Derow, M. R. (2017). The Aftermath of Romantic Betrayal: What's Love Got to Do with It? *Current Psychology*, *36*(3), 504–515. https://doi.org/10.1007/s12144-016-9438-y

Couch, L. L., & Olson, D. R. (2016). Loss Through Betrayal: An Analysis of Social Provision Changes and Psychological Reactions. *Journal of Loss and Trauma*, *21*(5), 372–383. https://doi.org/10.1080/15325024.2015.1108789

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

Doré, B., Ort, L., Braverman, O., & Ochsner, K. N. (2015). Sadness Shifts to Anxiety Over Time and Distance From the National Tragedy in Newtown, Connecticut. *Psychological Science*, *26*(4), 363–373. https://doi.org/10.1177/0956797614562218

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-

Sentencing Outcomes. *Psychological Science*, *17*(5), 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x
Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and How Implicit First Impressions Can Be Updated. *Current Directions in Psychological Science*, *28*(4), 331–336. https://doi.org/10.1177/0963721419835206
Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org/10.1037/0022-3514.38.6.889
Fourie, M. M., Hortensius, R., & Decety, J. (2020). Parsing the components of forgiveness: Psychological and neural mechanisms. *Neuroscience & Biobehavioral Reviews*, *112*, 437–451. https://doi.org/10.1016/j.neubiorev.2020.02.020
Gallagher, R. J., Stowell, E., Parker, A. G., & Foucault Welles, B. (2019). Reclaiming Stigmatized Narratives: The Networked Disclosure Landscape of #MeToo. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–30. https://doi.org/10.1145/3359198
Gannon, T. A., Rose, M. R., & Ward, T. (2008). A Descriptive Model of the Offense Process for Female Sexual Offenders. *Sexual Abuse*, *20*(3), 352–374. https://doi.org/10.1177/1079063208322495
Gantman, A. P., & Van Bavel, J. J. (2015). Moral Perception. *Trends in Cognitive Sciences*, *19*(11), 631–633. https://doi.org/10.1016/j.tics.2015.08.004
Goldenberg, A., & Gross, J. J. (2020). Digital Emotion Contagion. *Trends in Cognitive Sciences*, *24*(4), 316–328. https://doi.org/10.1016/j.tics.2020.01.009
Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology*, *48*(1), 250–256. https://doi.org/10.1016/j.jesp.2011.08.006
Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141
Heath, A., Fisher, S., & Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science*, *8*(1), 297–333. https://doi.org/10.1146/annurev.polisci.8.090203.103000
Higgins, E. T., & Brendl, C. M. (1995). Accessibility and Applicability: Some "Activation Rules" Influencing Judgment. *Journal of Experimental Social Psychology*, *31*(3), 218–243. https://doi.org/10.1006/jesp.1995.1011
Kachen, A., Krishen, A. S., Petrescu, M., Gill, R. D., & Peter, P. C. (2021). #MeToo, #MeThree, #MeFour: Twitter as community building across academic and corporate institutions. *Psychology & Marketing*, *38*(3), 455–469. https://doi.org/10.1002/mar.21442
Kantor, J., & Twohey, M. (2017). Harvey Weinstein Paid Off Sexual Harassment Accusers for Decades. *The New York Times*. https://www.nytimes.com/2017/10/05/us/harvey-weinstein-harassment-allegations.html
Kihlstrom, J. F. (2013). The person-situation interaction. In *The Oxford handbook of social cognition.* (pp. 786–805). Oxford University Press.
Klein, S. B., Sherman, J. W., & Loftus, J. (1996). The Role of Episodic and Semantic Memory in the Development of Trait Self-Knowledge. *Social Cognition*, *14*(4), 277–291. https://doi.org/10.1521/soco.1996.14.4.277

Kovács, G. (2020). Getting to Know Someone: Familiarity, Person Recognition, and Identification in the Human Brain. *Journal of Cognitive Neuroscience*, *32*(12), 2205–2225. https://doi.org/10.1162/jocn_a_01627

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

Maryn, A. G., & Dover, T. L. (2023). Who gets canceled? Twitter responses to gender-based violence allegations. *Psychology of Violence*, *13*(2), 117–126. https://doi.org/10.1037/vio0000436

McCullough, M. E. (2001). Forgiveness: Who Does It and How Do They Do It? *Current Directions in Psychological Science*, *10*(6), 194–197. https://doi.org/10.1111/1467-8721.00147

Mende-Siedlecki, P. (2018). Changing our minds: The neural bases of dynamic impression updating. *Current Opinion in Psychology*, *24*, 72–76. https://doi.org/10.1016/j.copsyc.2018.08.007

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013a). Diagnostic Value Underlies Asymmetric Updating of Impressions in the Morality and Ability Domains. *Journal of Neuroscience*, *33*(50), 19406–19415. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013b). Diagnostic Value Underlies Asymmetric Updating of Impressions in the Morality and Ability Domains. *Journal of Neuroscience*, *33*(50), 19406–19415. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Mende-Siedlecki, P., & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience*, *11*(9), 1489–1500. https://doi.org/10.1093/scan/nsw058

Metzler, H., Rimé, B., Pellert, M., Niederkrotenthaler, T., Di Natale, A., & Garcia, D. (2023). Collective emotions during the COVID-19 outbreak. *Emotion*, *23*(3), 844–858. https://doi.org/10.1037/emo0001111

Mohammad, S. M. (2016). Sentiment Analysis. In *Emotion Measurement* (pp. 201–237). Elsevier. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Murphy, S. C. (2017). A Hands-On Guide to Conducting Psychological Research on Twitter. *Social Psychological and Personality Science*, *8*(4), 396–412. https://doi.org/10.1177/1948550617697178

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv:1103.2903 [Cs]*. http://arxiv.org/abs/1103.2903

North, A., Grady, C., McGann, L., & Romano, A. (2020). 262 celebrities, politicians, CEOs, and others who have been accused of sexual misconduct since April 2017. *Vox*. https://www.vox.com/a/sexual-harassment-assault-allegations-list

Park, B., & Young, L. (2020). An association between biased impression updating and relationship facilitation: A behavioral and fMRI investigation. *Journal of Experimental Social Psychology*, *87*, 103916. https://doi.org/10.1016/j.jesp.2019.103916

Pennebaker, J. W. (1997). Writing About Emotional Experiences as a Therapeutic Process. *Psychological Science*, *8*(3), 162–166. https://doi.org/10.1111/j.1467-9280.1997.tb00403.x

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. 26.

Piazza, J., Sousa, P., Rottman, J., & Syropoulos, S. (2019). Which Appraisals Are Foundational to Moral Judgment? Harm, Injustice, and Beyond. *Social Psychological and Personality Science*, *10*(7), 903–913. https://doi.org/10.1177/1948550618801326

Pryor, J. B., LaVite, C. M., & Stoller, L. M. (1993). A Social Psychological Analysis of Sexual Harassment: The Person/Situation Interaction. *Journal of Vocational Behavior*, *42*(1), 68–83. https://doi.org/10.1006/jvbe.1993.1005

Rajapakse, T. (2020). *Simple Transformers* [Python]. https://simpletransformers.ai

Reeder, G. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology*, *44*(4), 736–745. https://doi.org/10.1037/0022-3514.44.4.736

Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, P. (1978). From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology*, *14*(3), 237–255. https://doi.org/10.1016/0022-1031(78)90013-6

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108 [Cs]*. http://arxiv.org/abs/1910.01108

Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020). Sarcasm detection using machine learning algorithms in Twitter: A systematic review. *International Journal of Market Research*, *62*(5), 578–598. https://doi.org/10.1177/1470785320921779

Schöne, J. P., Parkinson, B., & Goldenberg, A. (2021). Negativity Spreads More than Positivity on Twitter After Both Positive and Negative Political Situations. *Affective Science*, *2*(4), 379–390. https://doi.org/10.1007/s42761-021-00057-7

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*(2), 195–202. https://doi.org/10.1037/0022-3514.61.2.195

Sherman, J. W., & Bessenoff, G. R. (1999). Stereotypes as Source-Monitoring Cues: On the Interaction Between Episodic and Semantic Memory. *Psychological Science*, *10*(2), 106–110. https://doi.org/10.1111/1467-9280.00116

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756. https://doi.org/10.1038/s41562-018-0425-1

Simchon, A., Guntuku, S. C., Simhon, R., Ungar, L. H., Hassin, R. R., & Gilead, M. (2020). Political depression? A big-data, multimethod investigation of Americans' emotional response to the Trump presidency. *Journal of Experimental Psychology: General*, *149*(11), 2154–2168. https://doi.org/10.1037/xge0000767

Sisco, M. R., Bosetti, V., & Weber, E. U. (2017). When do extreme weather events generate attention to climate change? *Climatic Change*, *143*(1–2), 227–241. https://doi.org/10.1007/s10584-017-1984-2

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. https://doi.org/10.1073/pnas.1014345108

Sykora, M., Elayan, S., & Jackson, T. W. (2020). A qualitative analysis of sarcasm, irony and related #hashtags on Twitter. *Big Data & Society*, *7*(2), 205395172097273. https://doi.org/10.1177/2053951720972735

Tambe, A. (2018). Reckoning with the Silences of #MeToo. *Feminist Studies*, *44*(1), 197. https://doi.org/10.15767/feministstudies.44.1.0197

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 178–185. https://doi.org/10.1609/icwsm.v4i1.14009

Wagner, D. D., Chavez, R. S., & Broom, T. W. (2019). Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. *WIREs Cognitive Science*, *10*(1). https://doi.org/10.1002/wcs.1482

Xiong, Y., Cho, M., & Boatwright, B. (2019). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement. *Public Relations Review*, *45*(1), 10–23. https://doi.org/10.1016/j.pubrev.2018.10.014

Zacharias, C. (2018). *Twint* [Python]. https://github.com/twintproject/twint