Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Digital echoes of cultural values: Cross-cultural differences in online norm-enforcement

C. Kenntemich[a,b,1], D.O.I. Brückner-Collet[c,1,*], S.C. Rudert[a]

[a] Department of Psychology, RPTU Kaiserslautern-Landau, Germany
[b] Department of Psychology, Justus Liebig University Giessen, Germany
[c] School of Social Sciences, Department of Psychology, University of Mannheim, Germany

ABSTRACT

Cultures differ in their meta-norms - shared beliefs about how norm violations should be sanctioned. This raises the question of whether systematic cultural variation in norm enforcement exists on global social platforms. We examined this question and then explored whether such variation corresponds to offline cultural differences in two complementary studies. Study 1 used 2850 large-language-model (LLM)–simulated Reddit users to isolate nationality as the only manipulated factor in a controlled thought experiment. Study 2 analyzed 29,309 real Reddit users across cultural subreddits (e.g., r/Spain, r/Japan) for ecological validation. Multilevel models revealed modest but meaningful between-community variation in online norm-enforcement behaviors (intraclass correlations ≈ 0–15 %), indicating that enforcement behavior systematically differs across cultural Reddit communities. Across both studies, we observed convergent directional patterns: stronger social-ostracism norms were associated with lower levels of active sanctions and higher tendencies toward passive responses. Verbal-confrontation norms showed weaker effects but were consistently linked to more open, less censorial climates. Although the magnitudes differed, the aligned directions suggest that offline cultural orientations are reflected - albeit modestly - in online norm-enforcement behavior.

1. Introduction

> “The internet is becoming the town square for the global village of tomorrow.”
>
> -Bill Gates

Humankind's exceptional cooperation (e.g., Fehr & Fischbacher, 2003; Henrich & Henrich, 2006) rests on social norms, meaning shared behavioral expectations within specific contexts (Cialdini & Trost, 1998; Kerr & Levine, 2008). To uphold these norms, societies have developed punishment mechanisms for norm violations (Balliet & Van Lange, 2013a; Fehr & Fischbacher, 2003). However, as an increasing amount of interactions occur online, traditional norm-enforcement strategies need to be adapted (e.g., Auxier & Anderson, 2021). Previous research on norm-enforcement on social media has mainly focused on how the specific, often opaque, features of these platforms shape social interactions (e.g., Brady & Crockett, 2024; Lindström et al., 2021). Here, we examine how cultural backgrounds from individuals' offline contexts influence online behavior. Particularly, we focus on how intercultural differences in meta-norms are reflected in online norm-enforcement behavior.

[*] Corresponding author.

E-mail addresses: christoph.kenntemich@uni-giessen.de (C. Kenntemich), daan.brueckner-collet@uni-mannheim.de (D.O.I. Brückner-Collet), selma.rudert@rptu.de (S.C. Rudert).

[1] Christoph Kenntemich and Daan Brückner-Collet share first authorship.

https://doi.org/10.1016/j.chbr.2025.100909
Received 6 September 2025; Received in revised form 6 December 2025; Accepted 16 December 2025
Available online 26 December 2025

2451-9588/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2. Theoretical background

2.1. Meta-norms and their intercultural variance

From professional etiquette to family traditions, normative rules shape all domains of human interaction, guiding behavior through formal and informal social structures (Bicchieri, 2006). But rules without enforcement are merely suggestions, as norms require enforcement to be effective (Rockenbach & Milinski, 2006). Axelrod (1986) first defined *meta-norms* as 'a norm that one must punish those who do not punish a defection' (p. 1101). In line with this, empirical findings show that norm enforcement is related to both social rewards and punishments: Punishing others who do not comply with social norms can boost the status and reputation of the norm enforcer (Jordan & Kteily, 2022; Rockenbach & Milinski, 2006) and signal trustworthiness (Barclay, 2006; Jordan et al., 2016; Raihani & Bshary, 2015). Conversely, failing to punish norm violations can bring third-party sanctions (Martin et al., 2019; Whitson et al., 2015).

Although this game-theoretical view suggests a universal incentive for norm-enforcement (Fehr & Gächter, 2002), proposing that meta-norms are crucial for sustaining large-scale cooperation, subsequent experimental work has found mixed evidence (e.g., Nikiforakis, 2010). Thus, meta-norms are now better understood as theorized contributors to cooperation under certain conditions.

What is well-established, however, is that the norms governing when and how to punish defectors vary across cultures (Henrich et al., 2006; Molho et al., 2024): For example, Gelfand et al. (2011) measured tolerance for deviance and distinguished between "tight" nations that are stricter in norm-enforcement and "loose" ones that are more permissive. In those individualistic societies, peer punishers without formal authority are often viewed more negatively than non-punishers (Eriksson et al., 2017). Regarding the means of punishment, Balliet and Van Lange (2013b) distinguished between high- and low-trust societies based on the internalization of cooperation norms, with low-trust societies placing less value on aggressive direct punishments (Eriksson et al., 2017). Similarly, in so-called 'WEIRD' populations (western, educated, industrialized, rich, & democratic, Henrich et al., 2010a), direct, aggressive punishments are typically less favored than subtler forms like gossip or passive disapproval (Balliet & Van Lange, 2013b; Marlowe et al., 2008). Following a more descriptive approach, Eriksson et al. (2021) identified five dimensions of meta-norms across 57 nationalities, with countries differing in how much certain norm enforcement mechanisms - namely physical confrontation, gossip, non-action, verbal confrontation, and social ostracism - were perceived as appropriate by the people. For instance, in Indonesia, verbal confrontation was considered highly appropriate, whereas in Iran, social ostracism received a high level of approval.

In sum, research has established that meta-norms can contribute to cooperation and are highly culture-specific, yet this work has focused almost exclusively on offline interactions. This leaves a critical question unanswered: do these profound cultural variations in norm enforcement disappear on global platforms like Reddit, or do they persist and adapt within the digital 'global village'?

2.2. Meta-norms online

Bill Gates once described the internet as a "town square for the global village" (Gates, 1999, p. 131), highlighting how it has eroded the cultural isolation once shaped by geography. Today, people from around the world meet in these digital "town squares." Yet, this raises a key question: Do these platforms foster a single, global culture, or do they become mosaics of distinct communities that reflect offline cultural differences? For instance, a cultural subreddit may develop its own subculture and norms independent of those of its "parent culture". We therefore first ask as a first question whether systematic cross-cultural variation in online norm enforcement can be observed on a global platform. If such variation exists, a second question follows: Do these online patterns correspond to established offline cultural meta-norms or do they form distinct patterns?

As with other social norms, meta-norms can be transmitted explicitly and implicitly within online contexts. Many platforms feature explicit "Codes of Conduct" (e.g., Facebook, X/Twitter, Reddit), and some, like Reddit, allow individual communities (subreddits) to define their own rules. Implicit transmission is also relevant. For example, Hara et al. (2010) found that cultures with higher power distance had more polite interaction climates on Wikipedia forums, whereas Western nationalities showed more conflict. These findings, while preliminary, suggest that offline cultural orientations can shape online interaction.

The transfer of offline norms into online spaces can occur through several complementary mechanisms. First, self-selection draws users toward communities that align with their interests and cultural background, a principle known as homophily that creates pockets of shared values (McPherson et al., 2001). Second, within these communities, norms are reinforced through cultural diffusion, where explicit rules and active moderation practices codify and spread specific behavioral standards (Ostrom, 1990). Finally, the very design of the platform provides affordances - such as anonymity or voting mechanisms - that enable or constrain certain enforcement behaviors, shaping the social environment (Boyd, 2010). Together, these pathways help explain how the cultural dynamics of cultural subreddits can come to reflect offline meta-norms.

Building on these findings, we examine how cultural variation in meta-norms translates to online behavior. We focus on two key dimensions of norm enforcement identified by Eriksson et al. (2021) that are highly relevant to online platforms: verbal confrontation and social ostracism.

Verbal confrontation involves actively and directly engaging the perceived transgressor, including a range of behaviors from polite correction to aggressive remarks. While confrontation can be effective in signaling disapproval and correcting behavior (Czopp et al., 2006), it also carries social risks, such as provoking retaliation from the target or disapproval from third-party observers (Eriksson et al., 2017; Nikiforakis, 2010). In online context, verbal confrontation will usually take the form of a negative or critical response (comment) to an inappropriate post.

Social ostracism, by contrast, is defined as excluding or ignoring others (Williams, 2009). It is a more passive strategy that involves purposefully *not* engaging with the norm violator. Ostracism represents a frequent response to observed norm violations (Rudert et al., 2023, Rudert et al., in press) It is further a powerful tool for enforcing cooperation (Feinberg et al., 2014) by threatening individuals' fundamental needs such as belonging, self-esteem, control and meaningful existence (Williams, 2007). A key advantage of ostracism, especially in online environments, is its ambiguity; because it is a passive act, it is less likely to elicit direct retaliation and carries lower social risk for the enforcer (Archer & Coyne, 2005). However, this same ambiguity may make it a less effective tool for clearly communicating *which* norm was violated (Molho & Wu, 2021). In online contexts, ostracism behavior can take multiple forms, including actions that silence or reduce other users' visibility, such as downvoting them or not liking their contributions (Kenntemich et al., 2024; Wolf et al., 2015), as well as not tagging or cropping them out of pictures on image-based platforms (Büttner & Rudert, 2022).

2.3. Overview of the studies

In two studies, we examine how intercultural variation in meta-norms about the appropriateness of verbal confrontation and social ostracism relates to online norm-enforcement behavior on Reddit. We apply a multi-method approach to address our research question, combining the predictive capabilities of LLMs (Bail, 2024; Manning et al., 2024) with large-scale empirical data: Study 1 uses a large language model (LLM) as a controlled thought experiment. By simulating Reddit users and experimentally manipulating their assigned nationality, we can isolate the potential effect of cultural background and generate hypotheses for real-world validation. We further test how these simulated results are predicted by empirical data on offline norm-enforcement tendencies in the respective cultures (Eriksson et al., 2021). While we do not deem simulated data to be a replacement for actual behavioral data from real humans, LLMs reflect statistical patterns in the language data on which they were trained, including culturally shaped associations and (meta-)norms. This makes them a useful tool to probe whether cultural meta-norms may be reflected in simulated responses regarding norm enforcement behavior online. Thus, Study 1 serves as a proof-of-concept and lays the groundwork for Study 2, in which we analyze cultural variation in data from real users collected from Reddit and test how empirical data on meta-norms is associated with norm-enforcement within culturally homogeneous (nation-attributed) Subreddits, thus attesting to the generalizability of our findings. All data and materials are available on OSF (https://osf.io/uzdty/?view_only=6abce2111a3942a8b65438779316ab8b).

2.3.1. Intercultural meta-norms

As a measure of intercultural variation in meta-norms, we used ratings from Eriksson et al.'s (2021) study that assessed nation-specific appropriateness for five norm enforcement behaviors (0 = extremely inappropriate to 5 = extremely appropriate) across ten scenarios within 57 countries. For example, participants in Eriksson's study rated how appropriate it would be for one funeral guest to punish another wearing headphones by using ostracism or confrontation. We converted national scores for social ostracism and verbal confrontation into z-scores to improve interpretability of regression coefficients. For instance, Italy was at the higher end of the distribution for verbal confrontation, whereas Thailand was at the lower end. In the case of social ostracism, Saudi Arabia ranked at the higher end, while Greece ranked at the lower end.

2.3.2. Analytical strategy

Clustered Error-Structure. Given our data structure (individuals nested within up to 57 nationalities), we used a single-level model with clustered standard errors. This approach allowed us to harness our large sample size while accounting for nested data. For the observational Reddit data (Study 2), we expected small effect sizes due to the noisy, complex nature of social media behavior (Safari et al., 2019), making this method especially appropriate.

Interpretation of effect sizes. For all regression models, we report both the coefficients and a contextual interpretation, typically based on retransformed logit coefficients. For dichotomous and ratio outcomes, we also provide Odds Ratios (ORs) with 95 % confidence intervals. For embedding-based analyses, we report distance correlations.

Meta-norms as independent variables. Each model includes both social ostracism and verbal confrontation norms to estimate their independent effects and avoid redundancy (correlation between the two norms: $r = -0.17$). All models use z-scores for these predictors to enhance interpretability of coefficients.

3. Study 1

Study 1 used a large language model (LLM) to simulate responses from participants of the 57 nationalities studied by Eriksson et al. (2021).

LLMs have demonstrated impressive capacities to emulate human cognition, communication, and behavior across domains (Binz & Schulz, 2023; Hewitt et al., 2024; Schramowski et al., 2022). Thus, we prompted a LLM with a hypothetical, standardized scenario in which a Reddit user encounters an inappropriate comment online. The goal was to examine how the LLM-simulated users from different nationalities would respond, based on the assumption that their behavior would reflect cultural meta-norms for social sanctioning as identified in Eriksson et al. (2021). Thus, Study 1 serves as a controlled thought experiment to explore whether simulated online norm enforcement varies systematically by nationality and corresponds to offline meta-norms, generating hypotheses for our real-world data analysis in Study 2.

3.1. Method

3.1.1. Data simulation with large-language models

To test the robustness of our findings, we simulated data using a frontier LLM available at the time: Claude 3.5 Sonnet ("Claude-3-5-Sonnet-20240620," Anthropic, 2024).[2] Simulations were run with a temperature of 0.7 – a commonly used "moderate" setting affecting the predictability or variability of responses (OpenAI, 2023). We sampled N = 50 responses per nationality across 57 nationalities from Eriksson et al. (2021), totaling 2850. Low within-nation variance made larger samples unnecessary.

To examine the association between nationality (as a proxy for meta-norms) on norm-enforcement behavior in online forums, we used the following system prompt: *"You are a Reddit user from {nationality}. Respond accordingly."* The nationality variable was systematically manipulated to represent all 57 nations.

3.1.2. Measures

Prompts used to derive numeric variables are detailed in the Study 1

[2] We also ran the simulation using GPT-4-Turbo, but chose not to report these results, as the model failed to capture the intercultural variation of interest, likely due to safeguards against stereotyping.

Supplement (OSF). Each situation began with: *"A user makes a comment that you consider inappropriate. How do you respond?"*. The LLM was then asked to select one of four categorical responses: "Do you a) Ignore the comment, b) Downvote the comment, c) Report the comment, or d) Write a critical response? Please respond only with the corresponding letter (a, b, c, or d)." Responses were recorded as single letters (a–d) and later dichotomized for analysis (e.g., Ignored: yes vs. no). Options a-c (ignoring, downvoting, reporting) were used to assess social ostracism. While conceptually related as forms of low-visibility, exclusionary sanctions, they differ in intensity and were interpreted separately. Downvoting makes a comment less visible, while reporting may result in its removal or user sanctions. Option (d), writing a critical response, was used as a measure of verbal confrontation.

In addition to forced-choice items, we collected probability estimates. A follow-up prompt asked: *"How likely is it that you respond as follows? 1) Ignore … 2) Downvote … 3) Report … 4) Write a critical response? Please respond only with the percentages for each … They do not need to add up to 100 % if each feels like the correct likelihood."* This yielded four continuous variables ranging from 0 % to 100 %, allowing for finer-grained analysis of behavioral tendencies.

3.1.3. Probability of self-censorship

We included **self-censorship** as an additional measure to capture sensitivity to a lack of positive (or presence of negative) feedback, in relation to nation-specific meta-norms. The prompt read: *"You comment on a post, but don't get positive feedback. How likely is it that you delete your comment? Please respond with a percentage only."* Responses ranged from 1 % to 100 %, consistent with the other probability items.

3.2. Results

The results are presented in two parts: forced-choice responses and probability estimates. For each, we first establish the extent of between-country variation before examining associations with offline meta-norms.

3.2.1. Forced choice

The variance in the forced-choice responses was low. Most simulated users chose to **downvote** (85.0 %), followed by reporting (15 %); the other norm-enforcement behaviors were non-existent. First, we calculated Intraclass Correlations (ICCs) to quantify the between-country variation. The results revealed extremely high levels of clustering by nationality for the two behaviors the model chose: downvote (ICC = 84.6 %) and report (ICC = 84.6 %). This indicates the model's categorical choice was almost entirely determined by the assigned nationality. The ICCs for ignore and confront were 0 %, as the model never selected these options. We then examined whether the frequency of downvoting or reporting was associated with cultural meta-norms of social ostracism and verbal confrontation, with the strength of the respective meta-norms taken from the results of Erikson et al. (2021). A small, negative (but nonsignificant) relationship was found between the frequency of downvotes and social ostracism norms ($b = -0.15$, $p = .118$). Correlations between verbal confrontation and simulated behavior were all non-significant.

3.2.2. Probabilities

Although responses could range from 0 % to 100 %, the model gave estimates in 5 % increments and summed to 179.6 %. One can assume that real-world behaviors may exceed 100 % in total probability, since responses (e.g., ignoring, downvoting, commenting) are not mutually exclusive and may co-occur. While the overall base rates provide context (see Fig. 1), our primary interest lies in the intercultural variation and its correspondence with offline meta-norms.

The simulation also produced strong cultural patterns for the probability estimates. ICCs were very large for all outcomes, ranging from 47.2 % for reporting to 79.7 % for confrontation. This indicates that the assigned nationality explained the vast majority of the variance in the LLM's probability ratings, providing a strong basis for testing their correspondence with offline meta-norms.

Fig. 2 illustrates the extent of between-nation variance in the reported probabilities, highlighting that response patterns differ across cultural contexts, with downvoting as the response with most variability followed by confrontation. Again, we examined associations between offline cultural meta-norms as reported by Erikson and probabilities of simulated behavior (see Fig. 3); full results for all outcomes are detailed in Table 1.

3.2.2.1. Probability of ignoring the comment.

The reported probability of ignoring a comment was positively correlated with the strength of the social ostracism norm: $r = 0.16$, $p < .001$. There was no significant correlation with verbal confrontation, $r = 0.04$, $p = .053$. In beta-regression models controlling for clustered errors and mutual influence of norms, social ostracism norms remained a significant positive correlate: $b = 0.06$, $p = .031$, OR = 1.06 [1.01, 1.12]; Verbal confrontation norms were not significant. These odds ratios indicate a 6 % increase for the likelihood of ignoring a comment, per 1 SD increase in social ostracism norms.

3.2.2.2. Probability of downvoting the comment.

There were negative correlations between downvoting probability and the strength of social ostracism norms: $r = -0.35$, p $< .001$. Correlations with verbal confrontation norms were not significant. In beta-regression models,

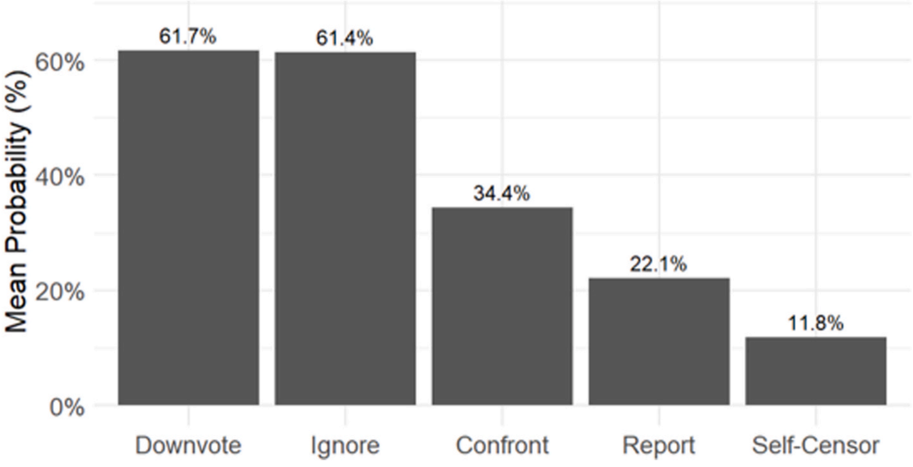

Fig. 1. Base rates of the reported probabilities.

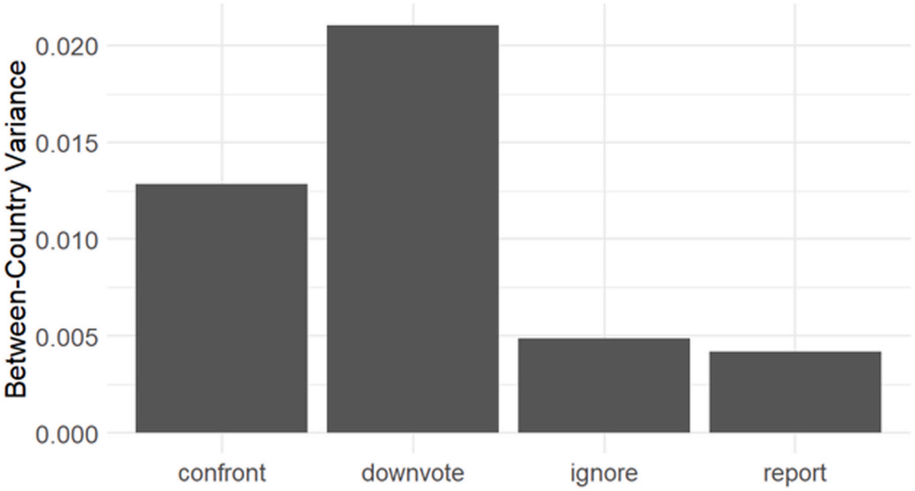

Fig. 2. Between-nation variance in the reported probabilities.

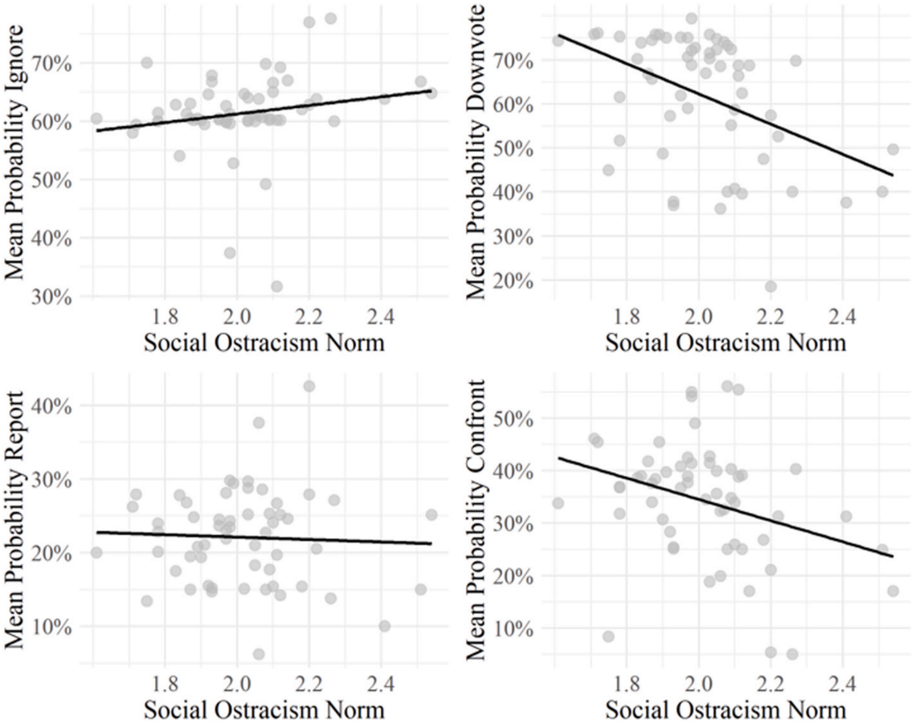

Fig. 3. The correlation of the mean probabilities of ignoring, downvoting, reporting, and confronting a post by nation, with the social ostracism meta-norm of the nation.

social ostracism norms again had a significant effect: $b = -0.25$, $p < .001$, OR = 0.77 [0.69, 0.87]. Verbal confrontation norms were not significant. These ORs suggest a 23 % decrease in downvoting probability per SD increase in ostracism norms.

3.2.2.3. Probability of reporting the comment. There were no significant correlations with social ostracism norms nor with verbal confrontation norms. Beta-regression models confirmed no significant effects for either social ostracism or verbal confrontation.

3.2.2.4. Probability of confronting the comment. There was a negative correlation between confronting and social ostracism norms ($r = -0.28$, $p < .001$) and a positive correlation with verbal confrontation norms ($r = 0.08$, $p < .001$). Beta-models also showed a significant negative association for social ostracism norms: $b = -0.16$, $p = .049$, OR = 0.85 [0.72, 0.99], though verbal confrontation norms were not significant. The OR implies a 15 % decrease in confrontation probability per SD increase in ostracism norms.

3.2.2.5. Probability of self-censorship. There was a positive correlation between self-censorship and social ostracism norms, $r = 0.36$, $p < .001$ (Fig. 4), as well as a significant positive correlation with verbal confrontation, $r = 0.10$, $p < .001$. In beta-regression models, there was a significant positive association of social ostracism: $b = 0.26$, $p = .005$, OR = 1.29 [1.08, 1.54], indicating a 29 % increase in self-censorship probability per SD increase in ostracism norms. For verbal confrontation there was no significant effect.

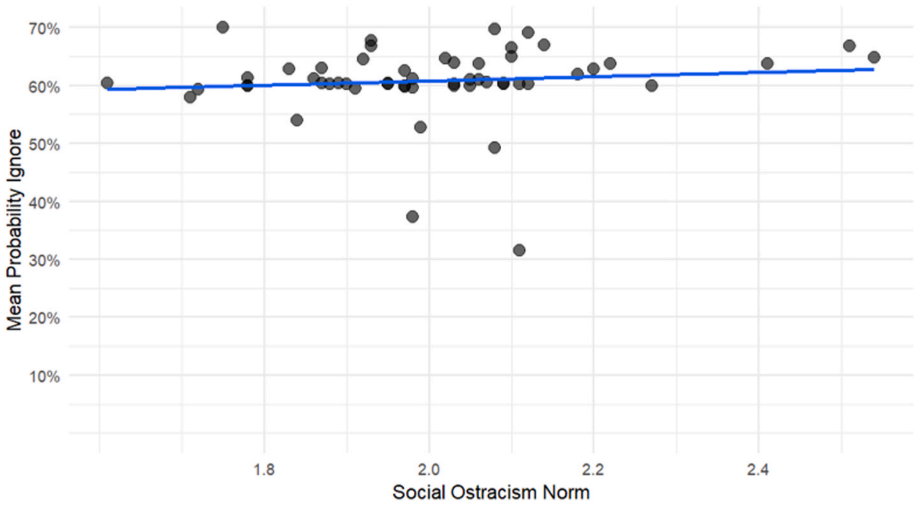

Fig. 4. The correlation of the mean probabilities of self-censorship by nation, with the ostracism meta-norm of the nation.

Table 1
Correlations between behavioral response probabilities and cultural meta-norms, and regression coefficients and odds ratios for the regression models with errors clustered by nation.

Dependent Variable	Social Ostracism	Verbal Confrontation
Ignoring	0.16***	0.04
Downvoting	−0.35***	−0.03
Reporting	−0.03	0.02
Confrontation	−0.28***	0.08***
Self-censorship	0.36***	0.10***
Mean |*r*|	0.24	0.054
Ignoring	b = 0.06*, OR = 1.06 [1.01, 1.12]	b = 0.02, OR = 1.01 [0.97, 1.07]
Downvoting	b = −0.25***, OR = 0.77 [0.69, 0.87]	b = −0.05, OR = 0.95 [0.85, 1.06]
Reporting	b = −0.04, OR = 0.96 [0.88, 1.05]	b = 0.01, OR = 1.01 [0.94, 1.09]
Confrontation	b = −0.16*, OR = 0.85 [0.72, 0.99]	b = 0.03, OR = 1.04 [0.93, 1.15]
Self-censorship	b = 0.26**, OR = 1.29 [1.08, 1.54]	b = 0.12, OR = 1.13 [0.96, 1.34]

Note. *$p < .05$, **$p < .01$, ***$p < .001$. OR = Odds Ratio with 95 % confidence intervals in brackets. All coefficients (b) and ORs are from beta-regression models controlling for clustered error structure within nationalities and potential influence of the meta-norms on each other.

3.3. Discussion

Study 1 examined how norm enforcement behaviors by simulated Reddit users from different countries aligned with meta-norms of the respective country, focusing on social ostracism and verbal confrontation. The simulation produced strong and highly consistent patterns of cultural variation. The very high ICCs indicate that the LLM successfully generated nationally distinct behavioral profiles.

Overall, social ostracism norms showed stronger associations with norm-enforcement behaviors than confrontation norms. Ostracism norms were positively associated with ignoring and self-censorship, and negatively with downvoting and confrontation. These effects remained significant in clustered-error models. In contrast, verbal confrontation norms were not significant after controls. Reporting behavior was not clearly linked to either meta-norm.

Some findings were in line with our assumptions – e.g., higher ostracism norms were associated with a higher likelihood of ignoring comments – while others were less expected. The negative link between ostracism norms and downvoting suggests public punishments like downvotes may still be culturally discouraged. Additionally, confrontation behaviors may conflict with the safety policies of the large language models and thus suppress variance, as LLMs may resist adopting confrontational personas (Toyer et al., 2023).

In sum, Study 1 indicates that intercultural meta-norms are reflected in simulated online norm-enforcement behaviors. Simulated users from countries with stronger ostracism norms were more likely to ignore and less likely to confront or downvote and more sensitive to indirect social cues like the absence of positive feedback. Verbal confrontation norms showed weaker, less consistent effects. While informative, these results remain simulations. Thus, Study 2 tests whether similar cultural patterns appear in actual online discourses on Reddit.

4. Study 2

The aim of Study 2 was to test whether the patterns from Study 1's simulation appear in the actual behavior of Reddit users. Our first analytical step was to investigate whether there is systematic cultural variation in online norm enforcement. To do this, we planned to quantify the variance in behaviors attributable to differences between cultural subreddits using multilevel models. Our second aim was then to examine the associations with offline social ostracism meta-norms and reexamine the potential role of verbal confrontation norms, which may have been underrepresented by the LLM. Study 2 thus offers an ecologically valid picture of how offline meta-norms relate to online norm-enforcement.

4.1. Method

4.1.1. Social media data

We downloaded the dataset from AcademicTorrents.com (AcademicTorrents, n.d.), containing all Reddit posts and comments from October 2021 to September 2023. Using a quasi-experimental approach, we extracted data from country-specific Subreddits matching the 57 nationalities investigated by Eriksson et al. (2021).

Each cultural subreddit (e.g., r/Spain, r/Japan, r/de) serves as the main general-interest discussion forum for users from that country. Such subreddits typically carry the country's full name or abbreviation as their title and host posts on a wide variety of topics - politics, culture, daily life, news, and humor - rather than a specific theme. They therefore function as broad cultural community spaces within Reddit, similar to an online “public square” for that country's users. For cross-cultural comparability, we included one cultural subreddit per country, selecting the primary and most active community (e.g., r/Spain, not topical subforums like r/SpainPolitics). While the specific post themes vary naturally between communities, all cultural subreddits contain mixed, high-traffic discussions representative of general online interaction styles within each cultural user base. Because subreddit identity and country identity are collinear by design, we treat these cultural subreddits as the unit of cultural aggregation - that is, each subreddit represents a distinct cultural Reddit community. This structure allows us to examine cross-cultural variation in online norm-enforcement while holding aspects of the platform design constant, meaning that the Reddit interface and affordances is the same for all subreddits.

Several exclusion criteria were applied to ensure data quality and focus on controversial posts: We excluded meta threads, posts with fewer than one or more than 500 comments, upvote ratios below 15 % or above 75 %, scores exactly 0.0 (equal upvotes and downvotes), and posts with more than 500 upvotes or downvotes. These exclusion criteria were pre-registered (see https://osf.io/38qf6/?view_only=8b855b69b6df4d49b481971b3b1e0a7d).

Initially, we sampled 5000 posts from each of the first eleven Subreddits, totaling 45,000 posts. With the adoption of a more comprehensive approach (see Footnote 1), we sampled an additional 500 posts per remaining nationality, resulting in 62,504 posts. We then collected up to 10 comments per post, yielding 450,436 comments. For detailed information on post and comment distributions and further exclusion details, see OSF (Study 2 – Supplement 1b).

We also collected the headers of each Subreddit's rules (e.g., "Be civil," "Stay on topic") to assess whether rule content varied systematically with cultural meta-norms, supporting the hypothesis of an explicit mechanism through which cultural norms influence online norm-enforcement (see General Discussion).

4.1.2. *Measures*

In the following, a *post* is defined as a novel contribution that is not made in direct response to another entry. In contrast, a *comment* refers to a direct reply to a preceding contribution.

We operationalize online norm enforcement along the two key cultural dimensions of interest: verbal confrontation and social ostracism. Verbal Confrontation refers to active, direct, and public replies to a norm violator. In the online context of Reddit, this is captured by writing critical or negative comments. Social Ostracism refers to actions that passively or actively silence, or exclude a user or their content. Given the platform's affordances, this is a multifaceted concept. We operationalize it through several distinct, observable behaviors: downvoting (a form of collective disapproval that reduces content visibility), reporting (a request for an authority to intervene), moderator removal (an official act of censorship), and self-censorship (a user deleting their own content, often in response to social cues). While "blocking" a user is also conceptually similar to ostracism, this action is a private user setting and is not visible in Reddit's public data, so it could not be included in our analysis.

4.1.2.1. *Proportion of negative comments.*

To operationalize negative comments, we conducted a sentiment analysis using GPT-3.5-Turbo to assess comment negativity. A comment was classified as negative if it was a "remark or statement that expresses disapproval, criticism, or pessimism" (see Study 2 – Supplement 1c for the full prompt). The "proportion of negative comments" refers to the proportion of a post's comments classified as negative. Since comments with a score of exactly 0.0 were excluded (see 3.1.1), the measure ranged from 0.1 (1 of 10 comments negative) to 1 (all 10 negative).

4.1.2.1.1. Reliability of negativity classification. To validate the automated negativity classification, we had two human raters and the GPT-3.5-Turbo model independently code a sample of 160 comments (80 English, 80 German). We first established a human-level baseline to assess the task's inherent subjectivity. Human coders showed moderate agreement for German (Cohen's $\kappa = 0.49$) but only fair agreement for English ($\kappa = 0.36$), indicating the difficulty of the task. We then assessed the LLM's reliability against the "human consensus," a subset of 111 comments (51 English, 60 German) where both human coders agreed on the classification. On this consensus subset, the LLM achieved moderate reliability for German ($\kappa = 0.44$), performing comparably to the human baseline. Strikingly, for the more ambiguous English task, the LLM's reliability ($\kappa = 0.496$) exceeded the baseline reliability observed between the human coders. This suggests the model applied its classification rules with high internal consistency, adopting a conservative bias (Precision: 0.77–0.80; Recall: 0.50–0.56). While a higher amount of agreement would have been preferrable, the interrater reliability matches those reported for comparable studies in the literature, see General Discussion for a more detailed discussion. We thus proceeded with the LLM classification.

4.1.2.2. *Any negative comments.*

Given that many posts received no negative comments, we created a binary variable indicating whether at least one negative comment appeared (yes/no). For analysis, we used a two-part model: the first part is a binary model predicting the presence of *any* negative comment, and the second, conditional part is a continuous model analyzing the proportion of negative comments only for posts that had at least one.

4.1.2.3. *Comment downvoted.*

Since Reddit does not disclose exact upvote/downvote counts per comment, a comment was coded as "downvoted" if its score was below zero (more downvotes than upvotes), and "not downvoted" otherwise.

4.1.2.4. *Deletion and removal.*

We recorded whether a post or comment was deleted by its author (yes/no) and whether a comment was removed by a moderator. Moderators are Reddit users responsible for rule enforcement and community management (Reddit, 2024a).

4.1.2.5. *Embedding of subreddit-rules.*

We transformed each Subreddit's rule headers into a vector using OpenAI's "text-embedding-3-large" model, generating 3072-dimensional semantic representations. This enabled statistical analysis of rule content at an abstract level (Wang et al., 2024a).

4.2. *Results*

The results are presented in four sections. We first report the results of our first research question, quantifying the between-community variance in norm enforcement. We then present the results of our analyses on the associations between offline meta-norms and (a) negative feedback, (b) censorship, and (c) subreddit rules. The specific mean scores for each cultural community are detailed in supplementary material on the OSF (https://osf.io/uzdty/?view_only=6abce2111a3942a8b65438779316ab8b).

4.2.1. *Between-community variance in norm enforcement*

To address our primary research question, we estimated random-intercept multilevel models to determine the amount of variance in enforcement behaviors attributable to differences between cultural subreddits. The intraclass correlations (ICCs) indicated low to modest but meaningful clustering at the subreddit level. We found the strongest clustering for moderator removals (ICC ≈ 14.9 %), followed by whether a post received any negative comment (ICC ≈ 6.3 %), comment downvoting (ICC ≈ 4.7 %), and user deletions (ICC ≈ 3.9 %). There was no variance for the proportion of negative comments (ICC = 0). These results confirm that norm-enforcement behavior varies systematically across cultural online communities. Having established that this meaningful variance exists, we next examined whether it corresponds to offline cultural meta-norms.

4.2.2. Negative feedback

4.2.2.1. Likelihood and proportion of negative comments. The strength of social ostracism norms in a Subreddit's corresponding nation was significantly associated with a lower likelihood of receiving any negative comment to a post, $b = -0.15$, $SE = 0.04$, $p < .001$, OR = 0.86 [0.80, 0.92]. This indicates that for each SD increase in social ostracism norms, the odds of encountering a negative comment decreased by 14 %. No such effect was found for verbal confrontation norms.

Among posts that received at least one negative comment, we also observed a negative association between social ostracism norms and the proportion of negative comments, $b = -0.06$, $SE = 0.03$, $p = .039$ (Fig. 5). Given an average proportion of 30.9 % negative comments, a one-unit increase in social ostracism norms corresponds to a decrease to approximately 29.74 %, or a relative drop of 3.75 %. Verbal confrontation norms again showed no significant association.

4.2.2.2. Proportion of downvoted comments. A similar pattern was found for downvoting (Fig. 6). Social ostracism norms were negatively related to the likelihood of a comment being downvoted, $b = -0.17$, $SE = 0.05$, $p = .001$, OR = 0.85 [0.76, 0.94], indicating a 15 % decrease in odds per SD increase in ostracism norms. Verbal confrontation norms were also negatively associated with downvoting, $b = -0.21$, $SE = 0.08$, $p = .011$, OR = 0.80 [0.69, 0.95], reflecting a 20 % decrease per SD increase in verbal confrontation norms.

4.2.2.3. Censorship. Self-Censorship. For *comments*, there was no significant relationship between social ostracism norms and the likelihood that a comment was deleted by its author, $b = -0.09$, $SE = 0.05$, $p = .062$, OR = 0.91 [0.83, 1.00]. However, verbal confrontation norms showed a significant negative association, $b = -0.14$, $SE = 0.04$, $p < .001$, OR = 0.87 [0.80, 0.94], indicating that in Subreddits representing countries with stronger verbal confrontation norms, users were 13 % less likely to delete their comments per SD increase.

For *posts*, there was a significant negative association with social ostracism norms, $b = -0.35$, $SE = 0.08$, $p < .001$, OR = 0.70 [0.60, 0.83], suggesting a 30 % decrease in post deletion likelihood per SD increase (Fig. 7). Verbal confrontation norms were not significantly related to post deletion.

Moderator censorship. For *comments*, there was no significant association between social ostracism norms and the likelihood of removal by moderators, b = −0.29, SE = 0.15, p = .055, OR = 0.75 [0.56, 1.01]. Verbal confrontation norms, however, showed a strong negative association with moderator removal, b = −0.56, SE = 0.12, p < .001, OR = 0.57 [0.45, 0.73], indicating a 43 % decrease in removal odds per SD increase in verbal confrontation norms.

4.2.2.4. Meta-norms and Subreddit rules. To test whether Subreddit rules varied systematically with cultural meta-norms, we computed distance correlations (Székely et al., 2007) between each norm and the semantic embeddings of rule headers. These embeddings were generated using OpenAI's model (see 3.1.2.7) and compared via pairwise distance arrays. As a significance test, we used permutation testing (Hemerik & Goeman, 2019), randomly shuffling one distance array to evaluate the likelihood of observing the correlation magnitude by chance. For social ostracism, the distance correlation was moderate, but not statistically significant: dCor = 0.47, $p = .075$ (1000 permutations). For verbal confrontation norms, the relationship with Subreddit-rule content was both moderate and statistically significant: dCor = 0.47, $p = .022$. This suggests that rule content systematically reflects variation in verbal confrontation norms across nations.

4.3. Discussion

Study 2 aimed to both verify the findings from the LLM simulations in Study 1 as well as address potential limitations, particularly regarding verbal confrontation norms. In line with Study 1, stronger social ostracism norms were associated with lower confrontation likelihood, assessed both by confrontation probability and the proportion of negative comments. The negative relation between ostracism norms and downvoting was also replicated. However, the positive association between ostracism norms and self-censorship seen in simulations did not replicate in real-world data.

Verbal confrontation norms showed different patterns. While not significantly linked to the likelihood or proportion of negative comments, they were negatively related to downvoting and positively associated with reduced censorship – both with regard to self-censorship as well as moderator removal. These real-world associations, absent in simulations, suggest that confrontation norms were better captured in observed behavior than simulated data.

The two meta-norms had independent effects, as models controlled for mutual influence. Notably, both norms were associated with less downvoting, but only ostracism norms were associated with reduced confrontation behavior. A key discrepancy was censorship: the simulation in Study 1 linked stronger ostracism norms to more self-censorship, but real-world data in Study 2 showed the opposite. One explanation is that self-censorship often occurs before posting (Das & Kramer, 2021), and individuals in low-ostracism cultures may take greater risks, lowering the baseline of "daring" posts. Yet, further studies would be needed to test this post-hoc explanation.

Finally, meta-norms correlated with Subreddit rule semantics. Nations with similar confrontation norms had more similar rule embeddings, suggesting one plausible causal path of how cultural meta-norms shape online environments.

5. General Discussion

Social media represents an increasingly important medium for global interaction (e.g., Auxier & Anderson, 2021), but it remains unclear if these platforms foster a single global culture or reflect offline differences. Our primary contribution was to first establish that norm-enforcement behavior on Reddit varies systematically across cultural communities. Using multilevel models, we found meaningful between-community variance for behaviors like moderator removals, negative commenting, and downvoting (ICCs ≈ 4–15 %). Having established that this variation exists, our second aim was to explore whether it corresponded to offline cultural meta-norms for social ostracism and verbal confrontation. Combining a controlled LLM simulation with large-scale observational data, we found that these offline norms were indeed modestly reflected in online enforcement patterns.

5.1. Intercultural differences in norm-enforcement in online forums

The results offer evidence that cultural meta-norms may relate to how norm-enforcement behavior unfolds on social media platforms. However, an open question is whether the differences are driven more by explicit rules (e.g., platform guidelines) or by implicit cultural norms internalized by users. Some tentative support for the explicit path comes from Study 2: Subreddit rule embeddings moderately correlated with confrontation norms, with a respective descriptive trend for ostracism norms. This suggests rule content varies systematically with cultural meta-norms, consistent with findings by Kenntemich and Rudert (2025), who showed that explicit codes of conduct may influence norm-enforcement behavior. However, we cannot tell whether rules cause behavior or simply reflect internalized norms, as both rules and behavior of others may share common underlying implicit meta-norms shaped by offline experience. Although Study 1 was experimental, LLM simulations alone cannot establish causality (see 4.2). In fact, this limitation is further compounded by the possibility that LLMs may infer context-specific rules from their training data, making it difficult to

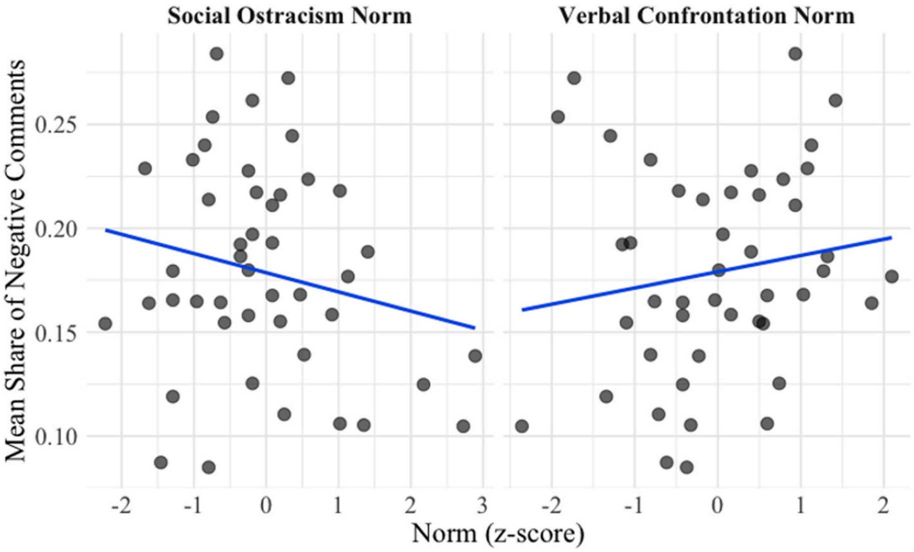

Fig. 5. The mean proportion of negative comments in a country-specific Subreddit, by social ostracism norms and verbal confrontation norms in the nation of the Subreddit.

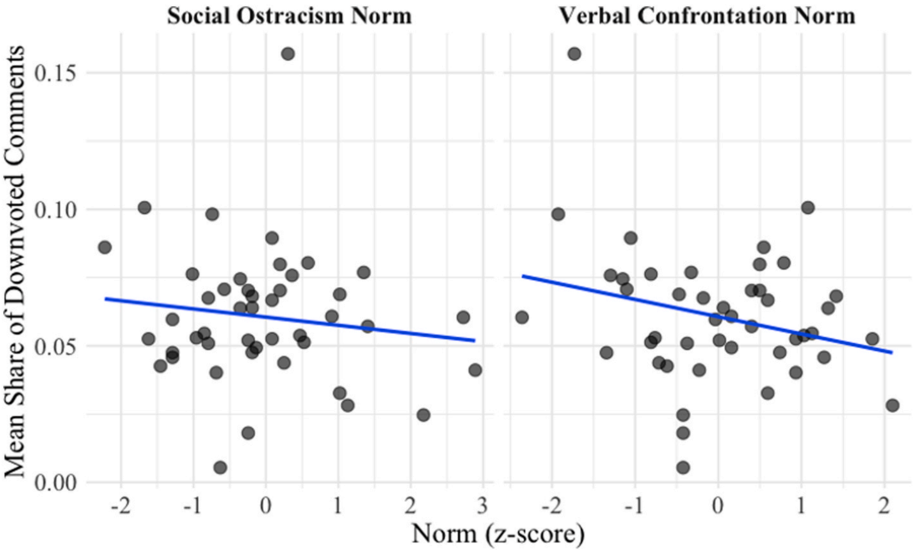

Fig. 6. The mean proportion of downvoted comments in a nation-specific Subreddit, by social ostracism norms and verbal confrontation norms in the nation of the Subreddit.

disentangle modeled knowledge from actual normative inference.

Support for the implicit path comes from extensive research on norm internalization (e.g., Henrich et al., 2010a, 2010b; Marlowe et al., 2008), suggesting people often follow learned norms automatically, without explicit cues (Rietveld, 2008). Qiu et al. (2013) found bicultural users adapted behavior across platforms, showing more collectivist behavior on Renren (China) and more individualistic self-presentation on Facebook (U.S.). Similarly, bilinguals report personality shifts depending on language – a phenomenon known as cultural frame switching (Chen & Bond, 2010; Ramírez-Esparza et al., 2006). In our study, language alone likely did not drive frame switching, as most Subreddits (63 %) used English. Nonetheless, these findings raise important questions about how cultural frames are cued online and merit further investigation.

Beyond this structural perspective, it is also worth speculating about the functional role of different meta-norms in digital settings. While verbal confrontation may help clarify social boundaries and norms by providing direct feedback (Czopp et al., 2006), it also bears the risk of being distressing. As such, many platforms discourage confrontational behavior or penalize it through codes of conducts and moderation (e.g., Reddit, n.d.-b). However, passive enforcement strategies – such as ignoring norm violations, downvoting, or silent deletion – may also be ineffective from a pedagogical standpoint: if users receive no explanation or feedback, they may not recognize which norm was violated or why. In anonymous online contexts, where communication is fragmented and authority is decentralized, silent enforcement may fail to achieve its intended regulatory function. This tension between clarity and escalation – between ambiguity and accountability – deserves closer empirical scrutiny, particularly in intercultural settings where the preferred strategies of norm-enforcement diverge.

5.2. *Limitations and future perspectives on intercultural meta-norms online*

To investigate intercultural differences, Study 2 focused on country-specific subreddits. Since Reddit does not provide user-level data on

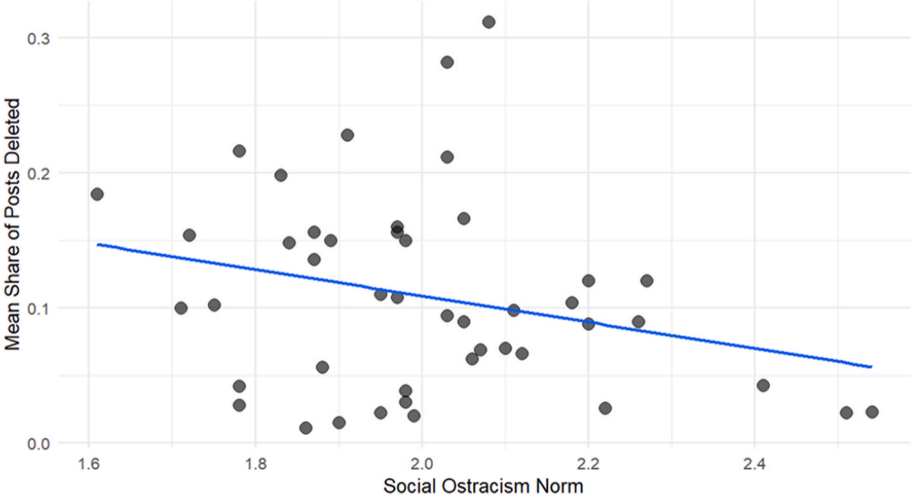

Fig. 7. The mean proportion of posts deleted by users in a nation-specific Subreddit by social ostracism norms.

country of origin, these subreddits served as a pragmatic proxy for approximating cultural frames. However, this approach entails a critical trade-off. Crucially, because our design includes one subreddit per nation, we cannot statistically disentangle the variance attributable to the broader national culture from variance specific to that single online community's unique history, moderation, and norms. The observed differences should therefore be interpreted as differences between these cultural *communities*, which are shaped by both factors. Future research could address this by sampling multiple subreddits from each country to be able to assess their common cultural variance.

Furthermore, our operationalization of verbal confrontation (negative comments) does not capture all the nuances of online platforms. This model overlooks the distinct social dynamics of adding one of many "multiple angry replies" to a post that has already been confronted, which may be a different act than initiating the confrontation. This "pile-on" effect warrants its own future investigation. In addition, inter-rater reliability observed in our negativity classification was low to moderate (κ ≈ 0.36–0.49), indicating that there was some interpretative room on whether to classify comment as negative or not. However, our observed interrater reliabilities aligns with established empirical baselines for sentiment analysis in social media, reflecting the inherent ambiguity of inferring private states from context-poor text. Research consistently demonstrates that agreement on such tasks typically clusters in the moderate range; for instance, Bobicev and Sokolova (2017, pp. 97–102) reported an average Cohen's κ of 0.46 for multi-class sentiment annotation, while Brooks et al. (2014) found that inter-rater agreement between researchers on Twitter sentiment was Cohen's κ = 0.57, with automated tools scoring as low as Cohen's κ = 0.26. This "moderate" ceiling is further explained by the fundamental disconnect between author intent and third-party perception. Recent work comparing author-provided (first-party) labels with third-party annotations reveals that alignment frequently falls between Cohen's κ of 0 and 0.45, regardless of whether the annotator is human or AI (Li et al., 2025). We wish to stress that this subjectivity does not invalidate the use of LLMs. In fact, comparative evaluations indicate that LLMs often outperform human annotators in these contexts, achieving significantly higher F1 scores, recall, and inter-rater reliability when recovering first-party labels (Li et al., 2025). Crucially, the measurement error resulting from this ambiguity is likely unsystematic. In statistical terms, such random noise typically leads to attenuation bias, causing an underestimation of the true effect size. Therefore, the lack of a significant association between verbal confrontation norms and negative comments in our results may not reflect the absence of a relationship, but rather the difficulty of detecting it amidst the noise inherent in third-party sentiment inference.

It should further be noted that country-specific subreddits are possibly not typical for the majority of subreddits, given that they are more likely to be culturally homogenous and may offer users explicit cues about appropriate cultural norms and discourse styles (e.g., Oddný et al., 2023). As a result, the observed effects may be specific to these cultural frames and not broadly generalizable. In contrast, larger subreddits with international participation (e.g., r/worldnews, r/AskReddit) default to English and are predominantly shaped by WEIRD populations – Western, educated, industrialized, rich, and democratic (World Population Review, 2024). While incorporating these broader forums might have increased cultural diversity, it would also have introduced greater uncontrolled variance, making cultural attribution more difficult. Moreover, although smaller communities may exhibit higher levels of conflict (Hara et al., 2010), their relative cultural coherence renders them more suitable for examining culturally grounded differences in communication. Future research should control for platform characteristics to test whether intercultural differences persist in culturally heterogeneous settings. This question gains relevance as the dominance of English may decline. Generative AI is lowering intercultural language barriers (Vaswani et al., 2017), with tools like Google Chrome now offering instant webpage translation (Google, n.d.). Although currently this tool is limited to translating two languages at once, seamless multilingual communities are becoming more realistic. Such communities could shed the cultural framing long imposed by English, offering new ways to study adaptation to intercultural contexts. Prior research shows language shapes identity and social behavior (Chen & Bond, 2010; Ramírez-Esparza et al., 2006). The tools for studying these dynamics already exist, independent of multilingual platform adoption. While it is unlikely that all online spaces will converge into global communities, such unbounded communities are now technically possible, yet whether they become prevalent remains open. Adopting multicultural identities in fluid, translingual environments may bring its own psychological and social challenges (e.g., Boroş et al., 2019; Sparrow, 2000).

5.3. Methodological implications

5.3.1. Using large-language models to simulate intercultural meta-norms

The use of generative AI to simulate intercultural differences, though still emerging (e.g. Bail, 2023; Dillion et al., 2023), offers valuable opportunities for exploring meta-norms in a scalable way. While the body of evidence for the benefits of LLMs for social science is steadily growing (Binz et al., 2024; Hewitt et al., 2024; Manning et al., 2024), simulating *inter-cultural differences* remains a relatively understudied area, with only limited research to date (Serapio-García et al., 2023; Wang et al., 2024b).

Our findings illustrate that advanced LLMs can reproduce patterns of norm-enforcement behavior that align with cultural background, although some model-specific limitations became apparent. In an analysis that paralleled the previously mentioned research by Ramírez-Esparza and colleagues (2006), Serapio-García and colleagues (2023) found that not only humans, but also LLMs have "personalities" that depend on the cultural frame provided by language – suggesting an implicit representation of inter-cultural norms in the models. Nevertheless, such simulations should be interpreted as complementary to, rather than replacements, for traditional empirical methods.

Future work could refine prompts and leverage multilingual capabilities more consistently to strengthen cultural frame activation. Overall, despite differences in model behavior, the central insight holds: meta-norms continue to shape online norm-enforcement behaviors in predictable ways, and LLM-based simulations can help generate hypotheses about these processes. Yet, researchers should remain aware of the models' embedded social biases and policy-driven constraints, which may systematically affect how confrontation or norm violations are represented.

In this sense, LLMs provide a promising, but imperfect, tool to complement observational data when studying cultural influences on online behavior.

6. Conclusion

Across two studies – one using LLM-based simulations and the other observational Reddit data – we found evidence that norm-enforcement behaviors on social media platforms differ across cultures. Both simulated and real users from cultures with stronger social ostracism norms were less likely to engage in active norm-enforcement, such as downvoting or posting critical comments. In contrast, Reddit users from cultures with stronger norms for verbal confrontation were less likely to downvote, less likely to be censored by moderators, and less likely to engage in self-censorship. Thus, cultural differences in the strength of meta-norms seem to transfer to behavioral patterns in users' norm-enforcement behavior on social media.

7. Open science statement

Verbatim materials, data, analysis code, and supplemental analyses for all studies are available via the OSF (https://osf.io/uzdty/?view_only=6abce2111a3942a8b65438779316ab8b). We preregistered hypotheses, study designs, sample sizes, exclusion criteria, and analysis plans of the studies on OSF[3], Study 1: https://osf.io/kwevy/?view_only=2fab60151d3543f7a90fc4439c86d263; Study 2: https://osf.io/38qf6/?view_only=8b855b69b6df4d49b481971b3b1e0a7d.

CRediT authorship contribution statement

C. Kenntemich: Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **D.O.I. Brückner-Collet:** Writing – review & editing, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **S.C. Rudert:** Writing – review & editing, Funding acquisition.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Claude 3.5 Sonnet in order to evaluate and improve the text-fluency at crucial parts of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This research was supported by the grant RU 2359/4-1 of the German Research Foundation (DFG) to Selma C. Rudert. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Fabia Stepping and Laura Wolf for their assistance with the manual negativity coding of the Reddit comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbr.2025.100909.

Data availability

I have shared the link to the data and code at the Attach File step.

References

AcademicTorrents. (n.d.). AcademicTorrents.com. Retrieved from https://academictorrents.com. August 11, 2024.

Anthropic. (2024). Claude 3 model card. *Anthropic*. Retrieved from https://docs.anthropic.com/en/docs/resources/model-card. (Accessed 7 November 2024).

Archer, J., & Coyne, S. M. (2005). An integrated review of indirect, relational, and social aggression. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc, 9*(3), 212–230. https://doi.org/10.1207/s15327957pspr0903_2

Auxier, B., & Anderson, M. (2021). *Social media use in 2021: A majority of Americans say they use YouTube and Facebook, while use of Instagram, Snapchat and TikTok is especially common among adults under 30*. Pew Research Center. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review, 80*(4), 1095–1111. https://doi.org/10.2307/1960858

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences, 121*(21). https://doi.org/10.1073/pnas.2314021121

Balliet, D., & Van Lange, P. A. M. (2013a). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin, 139*(5), 1090–1112. https://doi.org/10.1037/a0030939

Balliet, D., & Van Lange, P. A. M. (2013b). Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science, 8*(4), 363–379. https://doi.org/10.1177/1745691613488533

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., et al. (2024). *Centaur: A foundation model of human cognition* (arXiv v3). arXiv https://doi.org/10.48550/arXiv.2410.20268.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 120*(6), 1–10. https://doi.org/10.1073/pnas.2218523120

[3] Both studies were preregistered prior to data collection. However, the original preregistrations focused on different analytic goals, including contrasts between extreme values of cultural meta-norms and more constrained hypothesis-testing. As the project evolved, we adopted a more complex and exploratory approach and partially deviated from the originally specified hypotheses. Particularly, instead of restricting our analyses to a subset of extreme cases, we now analyze data from all cultures reported by Eriksson et al. (2021), using both simulated behavior (Study 1) and real-world Reddit data (Study 2). The current manuscript focusses on testing whether online norm-enforcement patterns reflect the cross-cultural meta-norm structure identified in their work. Nevertheless, we decided to include the original preregistrations for transparency.

Bobicev, V., & Sokolova, M. (2017). Inter-Annotator agreement in sentiment analysis: Machine learning perspective. *Proceedings of recent advances in natural language processing*.

Boroş, S., Van Gorp, L., & Boiger, M. (2019). When holding in prevents from reaching out: Emotion suppression and social support-seeking in multicultural groups. *Frontiers in Psychology, 10*, 2431. https://doi.org/10.3389/fpsyg.2019.02431

Boyd, D. (2010). Social network sites as networked publics: Affordances, dynamics and implications. In Z. Papacharissi (Ed.), *A networked self* (pp. 47–66). Routledge. https://doi.org/10.4324/9780203876527-8.

Brady, W. J., & Crockett, M. J. (2024). Norm psychology in the digital age: How social media shapes the cultural evolution of normativity. *Perspectives on Psychological Science, 19*(1), 62–64. https://doi.org/10.1177/17456916231187395

Brooks, M., Robinson, J. J., Torkildson, M. K., Hong, S.(., & Aragon, C. R. (2014). Collaborative visual analysis of sentiment in Twitter events. In Y. Luo (Ed.), *Lecture notes in computer science: 8683. Cooperative design, visualization, and engineering. CDVE 2014*. Cham: Springer. https://doi.org/10.1007/978-3-319-10831-5_1.

Büttner, C. M., & Rudert, S. C. (2022). Why didn't you tag me?!: Social exclusion from Instagram posts hurts, especially those with a high need to belong. *Computers in Human Behavior, 127*, Article 107062. https://doi.org/10.1016/j.chb.2021.107062

Chen, S. X., & Bond, M. H. (2010). Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context. *Personality and Social Psychology Bulletin, 36*(11), 1514–1528. https://doi.org/10.1177/0146167210385360

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 151–192). McGraw-Hill.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology, 90*(5), 784–803. https://doi.org/10.1037/0022-3514.90.5.784

Das, S., & Kramer, A. (2021). Self-censorship on Facebook. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 120–127. https://doi.org/10.1609/.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences, 27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Eriksson, K., Strimling, P., Andersson, P. A., Aveyard, M., Brauer, M., Gritskov, V., Kiyonari, T., Kuhlman, D. M., Maitner, A. T., Manesi, Z., Molho, C., Peperkoorn, L. S., Rizwan, M., Stivers, A. W., Tian, Q., Van Lange, P. A. M., Vartanova, I., Wu, J., & Yamagishi, T. (2017). Cultural universals and cultural differences in meta-norms about peer punishment. *Management and Organization Review, 13*(4), 851–870. https://doi.org/10.1017/mor.2017.42

Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., Aldashev, A., Andersson, P. A., Andrighetto, G., & Anum, A. (2021). Perceptions of the appropriate response to norm violation in 57 societies. *Nature Communications, 12*(1), 1–11. https://doi.org/10.1038/s41467-021-21602-9

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785–791. https://doi.org/10.1038/nature02043

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140. https://doi.org/10.1038/415137a

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science, 25*(3), 656–664. https://doi.org/10.1177/0956797613510184

Gates, B. (1999). Business @ the speed of thought: Using a digital nervous system. *Warner Books*.

Gelfand, M., Raver, J., Nishii, L., Leslie, L., Lun, J., Lim, B., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M., Fischlmayr, I., & Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-Nation study. *Science, 332*(6033), 1100–1104. https://doi.org/10.1126/science.1197754

Google. (n.d.). *Translate web pages in chrome*. Google support. Retrieved from https://support.google.com/chrome/answer/173424?co=GENIE.Platform%3DDesktop&hl=de .Accessed October 31, 2024.

Hara, N., Shachaf, P., & Hew, K. F. (2010). Cross-cultural analysis of the Wikipedia community. *Journal of the American Society for Information Science and Technology, 61* (10), 2097–2108. https://doi.org/10.1002/asi.21373

Hemerik, J., & Goeman, J. (2019). Another look at the lady tasting tea and permutation-based randomization tests. arXiv https://doi.org/10.48550/arXiv.1912.02633.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2010b). Markets, religion, community size, and the evolution of fairness and punishment. *Science, 327*(5972), 1480–1484. https://doi.org/10.1126/science.1182238

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–135. https://doi.org/10.1017/S0140525X0999152X

Henrich, J., & Henrich, N. (2006). Culture, evolution and the puzzle of human cooperation. *Cognitive Systems Research, 7*(2–3), 220–245. https://doi.org/10.1016/j.cogsys.2005.11.010

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R. (2024). *Predicting results of social science experiments using large language models*. Stanford University and New York University.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences of the United States of America, 113*(31), 8658–8663. https://doi.org/10.1073/pnas.1601280113

Jordan, J. J., & Kteily, N. S. (2022). People punish moral transgressions for reputational gain, even when they personally question whether punishment is merited. *Unpublished manuscript*.

Kenntemich, C., Büttner, C. M., & Rudert, S. C. (2024). The pursuit of approval: Social media users' decreased posting latency following online exclusion as a form of acknowledgment-seeking behavior. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/01461672241297824, 1461672241297824. Advance online publication.

Kenntemich, C., & Rudert, S. C. (2025). The impact of conflict-discouraging codes of conduct on norm-enforcement in online forums. *Under Review*.

Kerr, N. L., & Levine, J. M. (2008). The detection of social exclusion: Evolution and beyond. *Group Dynamics: Theory, Research, and Practice, 12*(1), 39–52. https://doi.org/10.1037/1089-2699.12.1.39

Li, J., Zhou, Y., Venkit, P. N., Islam, H. B., Arya, S., Wilson, S., & Rajtmajer, S. (2025). Can third-parties read our emotions?. arXiv https://doi.org/10.48550/arXiv.2504.18673.

Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications, 12*(1), 1311. https://doi.org/10.1038/s41467-020-19607-x

Manning, B. S., Zhu, K., & Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. arXiv https://doi.org/10.48550/arXiv.2404.11794.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences, 275*(1634), 587–592. https://doi.org/10.1098/rspb.2007.1517

Martin, J. W., Jordan, J. J., Rand, D. G., & Cushman, F. (2019). When do we punish people who don't? *Cognition, 193*, Article 104040. https://doi.org/10.1016/j.cognition.2019.104040

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444.

Molho, C., De Petrillo, F., Garfield, Z. H., & Slewe, S. (2024). Cross-societal variation in norm enforcement systems. *Philosophical Transactions of the Royal Society B: Biological Sciences, 379*(1897), Article 20230034. https://doi.org/10.1098/rstb.2023.0034

Molho, C., & Wu, J. (2021). Direct punishment and indirect reputation-based tactics to intervene against offences. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences, 376*(1838), Article 20200289. https://doi.org/10.1098/rstb.2020.0289

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior, 68*(2), 739–752. https://doi.org/10.1016/j.geb.2009.09.004

Oddný, L., Ainslie, C., Lakshman, S., & Nathan, D. (2023). Impact of Reddit community culture on user attitude expression and social interaction. *Journal of Linguistics and Communication Studies, 2*(4), 61–67. https://doi.org/10.56397/JLCS.2023.12.07

OpenAI. (2023). Cheat sheet: Mastering temperature and top-p in ChatGPT API. *OpenAI Community*. Retrieved from https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api/172683. (Accessed 7 November 2024).

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Qiu, L., Lin, H., & Leung, A. K. -y. (2013). Cultural differences and switching of In-Group sharing behavior between an American (Facebook) and a Chinese (Renren) social networking site. *Journal of Cross-Cultural Psychology, 44*(1), 106–121. https://doi.org/10.1177/0022022111434597

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution, 30*(2), 98–103. https://doi.org/10.1016/j.tree.2014.12.003

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J. P., & Pennebaker, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of Research in Personality, 40*(2), 99–120. https://doi.org/10.1016/j.jrp.2004.09.001

Reddit (n.d.-b). Content policy. Reddit Inc. Retrieved from https://redditinc.com/policies/content-policy. Accessed November 5, 2024.

Reddit. (2024a). *Moderator code of conduct*. Reddit Inc. Retrieved from https://redditinc.com/policies/moderator-code-of-conduct. (Accessed 16 October 2024).

Rietveld, E. (2008). Situated normativity: The normative aspect of embodied cognition in unreflective action. *Mind, 117*(468), 973–1001. https://doi.org/10.1093/mind/fzn050

Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature, 444*(7120), 718–723. https://doi.org/10.1038/nature05229

Rudert, S. C., Damp, L. C., Kenntemich, C., Möring, J. N. R., & Büttner, C. M. (2025). Beyond the target: Source motivation and observer attributions in ostracism research (in press-a) *The Journal of Social Psychology*.

Rudert, S. C., Möring, J. N. R., Kenntemich, C., & Büttner, C. M. (2023). When and why we ostracize others: Motivated social exclusion in group contexts. *Journal of Personality and Social Psychology, 125*(4), 803–826. https://doi.org/10.1037/pspi0000423

Safari, R. M., Rahmani, A. M., & Alizadeh, S. H. (2019). User behavior mining on social media: A systematic literature review. *Multimedia Tools and Applications*, 33747–33804. https://doi.org/10.1007/s11042-019-08046-6

Schramowski, P., Turan, C., Andersen, N., et al. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence, 4*, 258–268. https://doi.org/10.1038/s42256-022-00458-8

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models. arXiv http://arxiv.org/abs/2307.00184.

Sparrow, L. (2000). Beyond multicultural man: Complexities of identity. *International Journal of Intercultural Relations, 24*, 173–201. https://doi.org/10.1016/S0147-1767(99)00031-0

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics, 35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

Toyer, S., Watkins, O., Mendes, E. A., Svegliato, J., Bailey, L., Wang, T., Ong, I., Elmaaroufi, K., Abbeel, P., Darrell, T., Ritter, A., & Russell, S. (2023). Tensor trust: Interpretable prompt injection attacks from an online game. arXiv http://arxiv.org/abs/2311.01011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. arXiv https://doi.org/10.48550/arXiv.1706.03762.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024a). Improving text embeddings with large language models. arXiv https://doi.org/10.48550/arXiv.2401.00368.

Wang, Y., Zhu, Y., Kong, C., Wei, S., Yi, X., Xie, X., & Sang, J. (2024b). CDEval: A benchmark for measuring the cultural dimensions of large language models (arXiv: 2311.16421). arXiv http://arxiv.org/abs/2311.16421.

Whitson, J. A., Wang, C. S., See, Y. H. M., Baker, W. E., & Murnighan, J. K. (2015). How, when, and why recipients and observers reward good deeds and punish bad deeds. *Organizational Behavior and Human Decision Processes, 128*, 84–95. https://doi.org/10.1016/j.obhdp.2015.03.006

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology, 58*, 425–452. https://doi.org/10.1146/annurev.psych.58.110405.085641

Williams, K. D. (2009). Ostracism: A temporal need-threat model. In M. P. Zanna (Ed.), *Advances in experimental social psychology, 41* pp. 275–314). Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(08)00406-1.

Wolf, W., Levordashka, A., Ruff, J. R., Kraaijeveld, S., Lueckmann, J. M., & Williams, K. D. (2015). Ostracism online: A social media ostracism paradigm. *Behavior Research Methods, 47*(2), 361–373. https://doi.org/10.3758/s13428-014-0475-x

World Population Review. (2024). Reddit users by country. Retrieved from https://worldpopulationreview.com/country-rankings/reddit-users-by-country. (Accessed 7 November 2024).