

FACULTY OF PSYCHOLOGY

Department of Social, Economic, and Environmental Psychology

**Navigating Digital Norms:
Enforcement Mechanisms and User Adaptation
Strategies in Online Forums**

vom Promotionsausschuss des Fachbereichs 8: Psychologie

der Universität Kaiserslautern-Landau

zur Verleihung des akademischen Grades

Doktor der Philosophie (Dr. phil.) genehmigte

Dissertation

von Christoph Kenntemich, M. Sc.

Vorsitzende der Promotionskommission: Prof. Dr. Michaela Maier

Erste Berichterstatterin: Prof. Dr. Selma C. Rudert

Zweiter Berichterstatter: Prof. Dr. Rainer Greifeneder

Abgabe der Dissertation: 27.11.2024

Tag der Disputation: 10.03.2025

Acknowledgement

I first want to thank my supervisor, Selma Rudert, for providing the key ingredients for making my period of academic research a transformational experience of personal growth: trust and accompanying freedom to experiment. No project was too ambitious, independent of whether it matched my skillset at the time. We truly do grow with our challenges.

Next, I want to thank Christiane Büttner, for being the best academic sibling one could have asked for – and I’m typically not one for exaggerations. I think the phrase “being wise beyond your (academic) years”, probably captures it best; though it does neglect how generously you shared that wisdom. Thanks to Elianne, Melissa, and Rainer, Basel truly became my second academic home.

Out of the many people that accompanied my time in Landau, I want to thank my PhD-compatriots, from the first half - Tamara, Elena, and Caro – for showing me the ropes, or sometimes figuring them out together, as well as those from the second half – Helena, Lea, Larissa, and Lou – for plenty of stimulating discussions over lunch. Thanks also to Melanie, Claudia, and particularly Franziska and Petra for making the office days pleasantly sociable. Finally, Muireann and Sven – I wish circumstances (timing and distance, respectively) hadn't constrained our time together so much.

Außerdem möchte ich auch Line und besonders Alisa danken, dass sie meinem gelegentlich überschwänglichen Enthusiasmus für verschiedenste Themen stets standgehalten haben. Zuletzt möchte ich Oma, Opa, Mama und Papa danken, mit deren Unterstützung keine Herausforderung zu groß scheint.

Table of Content

1. Abstract	1
2. Introduction	2
2.1. <i>Mechanisms of norm-enforcement: Universals and intercultural differences</i>	3
2.2. <i>Norm-enforcement online</i>	5
3. Empirical insights on online norm-enforcement	7
3.1. <i>(How) do offline meta-norms manifest in online forums?.....</i>	7
3.2. <i>The explicit normative frame as a causal mechanism.....</i>	10
3.3. <i>Behavioral adaptations to norm-enforcement online: The case of social exclusion.....</i>	12
4. Discussion of the Present Insights and Perspectives for Future Research	13
4.1. <i>Norm-enforcement online: Not the town square of the global village... yet</i>	14
4.2. <i>Mechanisms of norm-enforcement in online forums: The origin of intercultural differences ..</i>	16
4.3. <i>New methodological approaches established in this work</i>	18
5. Conclusion	21
6. References.....	22
7. Appendices	33

1. Abstract

Norms and their enforcement are the life's blood of large-scale cooperation. Different cultures have landed at different sets of norms that structure their social togetherness, often shaped by cultural evolution within a set of geographical features. With the internet emerging as an increasingly influential medium of communication, a deep understanding of how norms emerge in online forums is increasingly important. This thesis investigates norm-enforcement in online forums, specifically focusing on 1) inter-culturally consistent parallels between offline- and online norm-enforcement patterns, 2) differences in explicit norms as a causal mechanism behind these parallels, and 3) behavioral consequences of online norm-enforcement. I find systematic differences in online norm-enforcement, that correlate with norms on offline norm-enforcement, with countries that are higher in social ostracism showing less active punishment on Reddit (i.e., negative comments and downvoting) and countries that are higher in verbal confrontation showing more negative comments, but less downvoting, as well as less top-down censorship through moderators. Furthermore, I find that the semantic content of the subreddits' codes of conduct also vary systematically, according to cultural norms on social ostracism and verbal confrontation. Furthermore, in a set of experimental studies, I find a causal impact of codes of conduct on norm-enforcement behavior, underlining the explicit codes of conduct as a likely causal path of intercultural differences in norm-enforcement behavior in online forums. Lastly, I investigate the behavioral consequences of one norm-enforcement behavior – social ostracism. Across different online forums, users show increased attention seeking following contributions that receive less positive attention. Implications of the set of projects, as well as their innovative methodology, are discussed and potential future research is outlined.

2. Introduction

Human societies exhibit an unprecedented scale of cooperation (Fehr & Fischbacher, 2003; Henrich & Henrich, 2006), underpinned by social norms - shared behavioral expectations that guide actions within specific contexts (Cialdini & Trost, 1998; Kerr & Levine, 2008). These norms function as a powerful social technology for coordinating large-scale interaction, with their effectiveness maintained through enforcement mechanisms, particularly the punishment of norm violations (Balliet & Van Lange, 2013; Fehr & Fischbacher, 2003). These fine-tuned mechanisms of social interaction (Gintis, 2003), face a paradigm shift, as online forums are becoming an increasingly dominant medium of social exchange (Auxier & Anderson, 2021; Dixon, 2022), with drastically differing conditions for transmitting socially relevant information (Lapidot-Leffler & Barak, 2012, 2015). Understanding how these altered conditions affect social regulation is crucial. Here, I want to investigate what determines norm-enforcement processes in online forums, and how this affects the targets of the norm-enforcement. I will first outline previous research on norm-enforcement generally, understanding the mechanisms involved and outlining relevant norm-enforcement mechanisms. Then, I introduce previous research that has been done on norm-enforcement online thus far, outlining how it has mainly been focused on how it has been shaped by the medium itself, with little focus on inter-medial patterns between offline- and online interaction. Within three projects, I then investigate 1) whether and how offline norm-enforcement behavior correlates with online norm-enforcement behavior across cultures, 2) how this relationship might be mediated by online “codes of conduct”, and 3) how social exclusion online, as a particularly prevalent norm-enforcement strategy in western countries, affects its targets. Lastly, I will conclude by integrating the findings and discussing relevant follow-up questions for future studies.

2.1. Mechanisms of norm-enforcement: Universals and intercultural differences

While norms themselves provide a rule-like framework for expected behavior in social settings (Cialdini & Trost, 1998), they remain mere suggestions without appropriate motivation to respect them (Rockenbach & Millinski, 2006). The effectiveness of norms in shaping behavior beyond individual self-interest thus critically depends on the right balance of incentives (Axelrod, 1986; Fehr & Fischbacher, 2003). While norms determine what behavior is considered appropriate (Bicchieri, 2005), meta-norms ensure incentives against inappropriate behavior (Strimling & Eriksson, 2014; Axelrod, 1986) by 1) designating enforcement responsibilities and 2) establishing legitimate forms of punishment. To assign responsibilities, meta-norms provide their own set of (dis-)incentives: Enforcers can gain status and enhanced reputation (Jordan & Ktelly, 2022; Rockenbach & Millinski, 2006), signal their trustworthiness to others (Barclay, 2006; Jordan et al., 2016; Raihani & Bshary, 2015), and avoid the potential costs of non-intervention (Martin et al., 2019; Whitson et al., 2015). Who is expected to take on the responsibility of norm-enforcement, according to meta-norms, can broadly be categorized in two groups: decentralized *peer-to-peer* norm-enforcement and centralized *top-down* norm-enforcement (Ostrom, 1990). Legitimate forms of punishment can vary from verbal or physical confrontation of deviants (Czopp et al., 2006; Swim et al., 1999), social ostracism of deviants (e.g., Hales et al., 2017), or gossip (e.g., Feinberg et al., 2014), providing different levels of incentives for adherence to norms.

While these described components of meta-norms – *when* to expect *what* kind of punishment, from *whom* - represent universal building blocks, cultural evolution has assembled them into remarkably different social architectures (Molho et al., 2024), across geographical (and historical) idiosyncrasies (Henrich et al., 2010b). First, there are quantitative differences in norm-enforcement – when to expect punishment. Gelfand and colleagues (2011) classified different countries along a spectrum of reactivity, measuring tolerance for deviance and strength

of punishments in 33 countries. In Gelfand's classification "tight" countries displayed heightened reactivity, with stronger punishments for lesser deviations, and "loose" countries were more permissive. Second, there are qualitative differences in the norm-enforcement across cultures – what kind of punishment to expect. In a comprehensive study on meta-norms in 57 societies, Eriksson and colleagues (2017) explicitly distinguished country-specific preferences in exact norm-enforcement behaviors, mapping the intercultural spectrum of opinions towards social ostracism, gossip, verbal confrontation, and even physical confrontation. Different strokes for different folks, indeed.

Lastly, there are differences in what social actors are considered responsible for norm-enforcement – from whom to expect punishment. The dominant culture in research on norm-enforcement – as for most social scientific research - has without a doubt been the WEIRD culture, an acronym for **W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic (Henrich, 2010a). Consequently, it was long debated to potentially be a universal human instinct for individuals to feel responsible for upholding norms of fairness and reciprocity (e.g., Nowak & Sigmund, 2000), through decentralized peer-to-peer punishment. But this underestimated the plasticity of behavioral norms, as high levels of altruistic peer-to-peer punishment rather proved to be a staple of large-scale societies (Marlowe et al., 2008), particularly those with high levels of impersonal trust (Balliet & Van Lange, 2013b) – both key-characteristics of WEIRD societies (Henrich et al., 2010b). Though WEIRD societies are conventionally categorized as individualistic (Hofstede, 2011), their commitment to universal principles of fairness and reciprocity (Balliet & Van Lange, 2013)—often at the expense of immediate material gain (Henrich et al., 2005)—highlight an underlying *collectivist mechanism*, implicitly guided by norms (Marlowe et al., 2008; Henrich et al., 2010b). In his highly influential paper on cooperation, Axelrod (1986) had already outlined the arguments for the game-theoretical effectiveness of such an individualistic responsibility to upholding norms, arguing for the effectiveness of '*a norm that one must punish those who do not punish a defection*' (Axelrod,

1986, p. 1101). Under the resulting omnipresent checks and balances, it is unsurprising that many researchers point toward this norm of shared responsibility for punishment as the fundament of the high levels of cooperation in WEIRD societies (Henrich et al., 2010b; Marlowe et al., 2008; Balliet & Van Lange, 2013).

Norm-enforcement is thus the binding force that maintains order in social environments from groups to societies. There are no universal meta-norms, guiding principles for when, how and by whom norms should be enforced, but such meta-norms are rather the result of cultural evolution and influenced by a culture's unique geographical and historical circumstances. It is thus crucial to investigate how patterns of norm-enforcement translate to the online world, where geographical circumstances are (for the most part) minimized, and the channels of communication differ compared to offline social environments.

2.2. Norm-enforcement online

While the exact channels of communication can vary between online forums (e.g., “Likes” on Facebook vs. “Upvotes” on Reddit), they have many similar components. The first general component is what has been defined as *paralinguistic digital affordances* (PDAs - Hayes et al., 2016): tools to communicate explicit social cues on social media, including “Favorites” or “Retweets” on X/Twitter, or Facebook’s “Like”-Button. Thus far there is little research how exactly PDAs might be used might be actively used for punishment. That they generally can be an effective tool for norm-enforcement has been shown by research, suggesting that a lack of positive PDAs might be a sufficient to induce negative emotions (Wolf et al., 2015; Hayes et al., 2018; also see 3.3.). Second, users can use comments to communicate discontent (Brady et al., 2021; Rost et al., 2016) – what I will henceforth call *confrontation*.

While both paralinguistic digital affordances and confrontation are both channels for decentralized peer-to-peer norm-enforcement, there are typically also centralized top-down channels of content moderation (UNESCO, 2014). While generally more centralized than the

peer-to-peer channels, the level of centralization varies between platforms, with Reddit - which is organized in forums ordered by topic, called *subreddits* – having decentralized moderators for each subreddit. The moderators’ task is generally to enforce the code of conduct, which can be considered as the ought-norms (i.e., injunctive interaction norms – Cialdini & Trost, 1998), for how people *should* and *should not* behave online (e.g., Reddit, 2024).

While there are obvious patterns of how online channels of communication can mirror offline channels of communication – e.g., Facebooks famous “Like” button is literally a “thumbs up” - it is nonetheless a vastly different medium of communication (Hayes et al., 2016; Lapidot-Leifler & Barak, 2012, 2015). It is therefore nonobvious whether intercultural differences in online norm-enforcement translate to online forums, and if they do, how this translation might manifest. What little research has been done, regarding this question, provides no conclusive evidence. Research by Hara et al. (2010) found that cultures with higher power distance (a Hofstede et al. 2011 measure that positively correlates with Gelfand's tightness score, $r = 0.461$, $p < .05$ – calculated from open-source data from the two papers) exhibited more polite interactions in Wikipedia creator forums. They also observed that Western countries tended toward more conflict and disagreement—a pattern generally aligning with the shared responsibility for norm-enforcement described by Balliet and van Lange (2013b). However, these findings should be considered preliminary, as the authors themselves emphasize the need for additional research to reach more definitive conclusions. Furthermore, even these preliminary findings do not aim at investigating the causal mechanism of how potential differences between cultures might translate – i.e., what is the causal mechanism that leads from intercultural differences in meta-norms to intercultural differences in norm-enforcement in online forums. Lastly, more research is needed to investigate the effectiveness of translated norm-enforcement behaviors. While previous research has started to investigate negative emotional consequences to rebuffs online (Hayes et al., 2016, 2018; Wolf et al., 2015),

behavioral change is the ultimate metric to capture the effectiveness of norm-enforcement (Fehr & Gächter, 2002; Chudek & Henrich, 2011).

3. Empirical insights on online norm-enforcement

In the following, I will outline three studies that contribute to understanding of norm-enforcement online, specifically focusing on 1) how offline norms on appropriate norm-enforcement (i.e., meta-norms) manifest in online forums, 2) the explicit normative frame of “codes of conduct” as a relevant causal mechanism, and 3) the effects of social ostracism, likely the most frequent form of norm-enforcement in online forums, on its targets.

3.1. (How) do offline meta-norms manifest in online forums?

Research on norm-enforcement online has, for the most part been focused on the how the medium itself shapes social interaction (e.g., Brady & Crockett, 2024; Lindström et al., 2021), but less focus on how norm-enforcement in online forums might be shaped by offline meta-norms. In a study on inter-cultural differences within Wikipedia creator forums, Hara and colleagues (2010) state a cautious first claim of inter-cultural differences in conflict behavior, as western creator forums – from countries with flatter hierarchies (Hofstede et al., 2011) – had a tendency towards more conflict and disagreement. Yet, even for this one-dimensional correspondence between offline norms and online conflict, the authors suggested that further research would be needed before any conclusions could be made.

In a set of two studies, Kenntemich, Brückner-Collet, & Rudert (2024) thus set out to achieve a more comprehensive overview of how offline meta-norms might be related to distinct online norm-enforcement practices. For this, they leveraged empirical work on intercultural variances in offline meta-norms, to investigate potential corresponding differences in the norm-enforcement behavior of culturally diverse social media users. Study 1 was a simulation study based on Large-Language Models (Hewitt et al., 2024; Manning et al., 2024), where Reddit-

users were simulated with an experimental manipulation of the nationality. As, to our knowledge, there had been no studies the aimed at simulating inter-cultural differences in social behavior with LLMs, we conducted a pre-study, in which we assessed the validity of the approach, by replicating the responses from a large-scale study on inter-cultural differences in meta-norms, by Eriksson and colleagues (2021). In short, we found that one of the two models was much better at capturing the inter-cultural variance (Claude 3.5 Sonnet outperformed GPT-4-Turbo), and that the inter-cultural differences for one of the two meta-norms were captured much better (social ostracism norms were replicated more accurately than verbal confrontation norms).

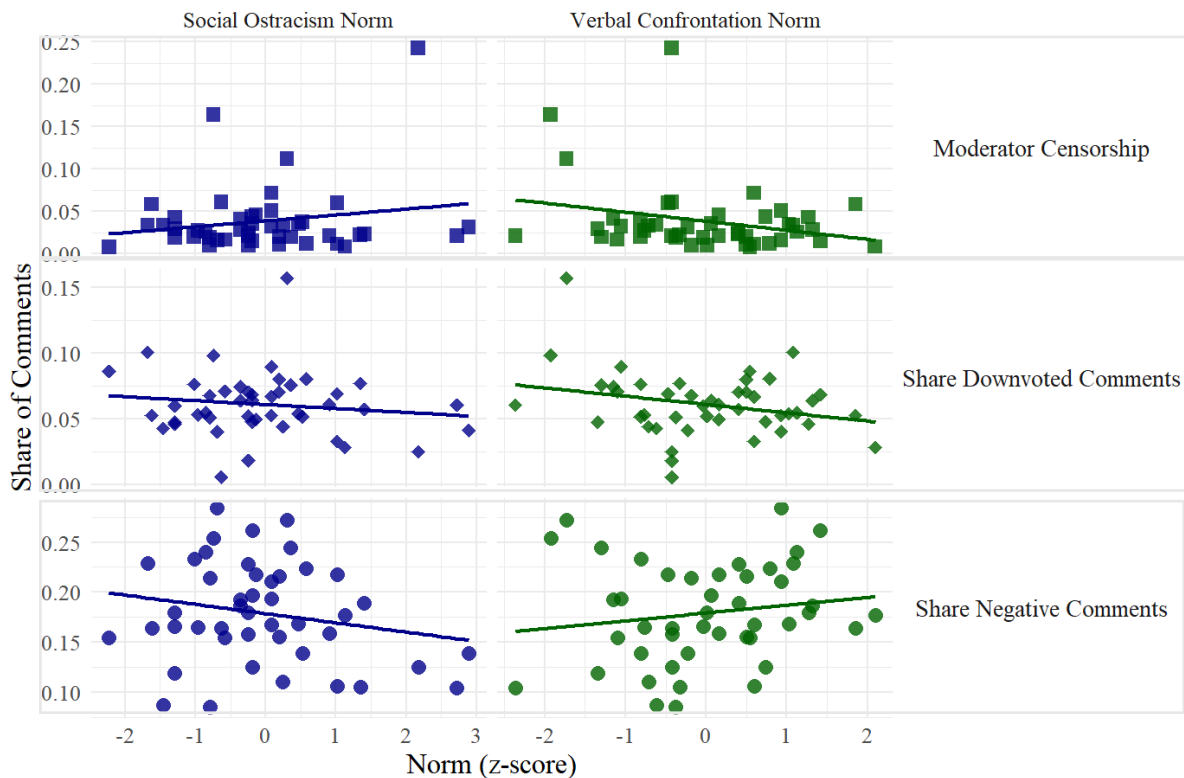
The simulated Reddit-users exhibited systematic differences across nationalities, in how they would respond to an inappropriate comment. These differences corresponded with offline meta-norms as they had been captured by Eriksson et al. (2021): Simulated users from countries that are more favorable towards social exclusion as a means of norm-enforcement less likely to respond actively to an observed norm violation: They reported a lower assumed probability of *confronting* the comment or, perhaps more surprisingly, *downvoting* the comment, while reporting a higher probability on *ignoring* the comment. Simulated users from countries that are more favorable to verbal confrontation showed fewer such systematic differences (perhaps unsurprisingly, following the validation-study) – with only a slightly greater tendency in the reported likelihood of *confronting* the comment.

The second study aimed at finding similar patterns in real Reddit data. In a quasi-experimental design, we investigated systematic differences in the norm-enforcement behavior across 45 culturally homogenous “Subreddits” (i.e., sub-forums specified by topics). Here, country-specific subreddits on countries that are more favorable towards social exclusion showed similar patterns to the simulated participants: They had a lower share of negative responses under controversial posts, and fewer downvoted comments. For countries that are

more favorable towards social exclusion we found more correlations than in the simulation study: Again, they were more likely to confront other users, and additionally, they were less likely to downvote, and there was less top-down censorship through moderation.

Figure 1

Association of norms on social ostracism (left) and verbal confrontation (right) with subreddit-based means of the share of comments that are removed by moderators (Moderator Censorship), the share of comments that had a negative score (Share Downvoted Comments), and the share of comments with a negative sentiment (Share Negative Comments), in Study 2 of Kenntemich, Brückner-Collet and Rudert (2024).



Note. “Share Downvoted Comments” lists the share of comments with a negative score (more downvotes than upvotes), because there are no specific data on the number of downvotes per comment, making comments with a negative score the only ones that were downvoted with certainty.

In sum, the studies show a systematic relationship between meta-norms offline and norm-enforcement practices online. This indicates that there are no general patterns of how

norms are enforced online, rather, the preferred strategies for online norm-enforcement correspond to favored strategies for offline norm-enforcement.

3.2. The explicit normative frame as a causal mechanism

Having established corresponding patterns between offline meta-norms and online norm-enforcement behavior, the question remains how these differences manifested: What is the causal path between offline differences in meta-norms and online differences in norm-enforcement behavior? One possibility would be, if internet forums with different dominant cultures differed in their *explicit* normative frame: In other words, interculturally varying forums have different *codes of conduct* - i.e., explicit rules on how to behave in a forum.

One indicator in favor of this causal path was an analysis from the study by Kenntemich, Brückner-Collet, & Rudert (2024), that investigated whether the codes of conduct in the different country-specific Subreddits differed systematically according to the country specific meta-norms. For this aim, we encoded the semantic content of the subreddit rules with *vector embeddings*, transforming the content of the rules into a high-dimensional (here 3,072 dimensions – see OpenAI, n.d.) numeric representation of its semantic meaning (see Wang et al., 2024), allowing to encode text data in a statistically processible form. Distance correlations between these vectors and the meta-norms revealed medium-sized correlations between the vector embeddings of countries' Subreddit's code of conduct and the countries' meta-norms. While this could indicate that the differences in online norm-enforcement between countries is caused by their different codes of conduct, it could also be, that the meta-norms are a shared covariate that determines both norm-enforcement behavior and the explicit rules, without them influencing each other. Yet, multiple studies on the impact of codes of conduct on social interaction undermine this explanation. In a large scale quasi-experimental study, Gibson (2019) showed that two Subreddits that were generally concerned with the same topic (“r/lgbt” and “r/ainbow”), but had drastically different codes of conduct, one being a “safe space” with

strict rules against offensive content, and the other advocating “free speech”, and therefore looser content moderation. In the “safe space” – r/lgbt, there was more top-down intervention and more self-censorship, compared to the “free speech” – r/ainbow. But the Subreddits did not only differ in their codes of conduct, so still, causal impact is still uncertain.

Consequently, Kenntemich & Rudert (2024) conducted a set of that investigated the causal impact of codes of conduct on norm-enforcement in online forums, specifically focusing on social ostracism and confrontation of a norm-violating target. In a set of 5 studies, we established a paradigm that simulated an online forum, to 1) allow for consistency between participants, 2) allow for both social ostracism and confrontation of the norm-violating target, and 3) allowed for the experimental variation of codes of conduct. Parallel to Gibson (2019), we manipulated the restrictiveness of the codes of conduct, with one encouraging conflict (i.e., “Every opinion is subjective and should be challenged”) and one discouraging conflict (i.e., “Every opinion is valid and should be accepted!”). Across three studies, we found that the rate of confrontation of norm-violating targets increased under norms that encouraged conflict, while the rate of social ostracism increased when the norm was conflict discouraging. Kenntemich & Rudert (2024) thus provides evidence that codes of conduct in online forums causally influence the chosen norm enforcement strategies.

Considering how semantic content of a countries’ Subreddit-rules correlate with countries’ social ostracism and verbal confrontation meta-norms, combined with the evidence on the causal impact of the codes of conduct (cf. Subreddit-rules), it seems likely that the previously outlined explicit causal path is at least part of the explanation for the correspondence between offline meta-norms and online norm-enforcement behavior, between countries. Nonetheless, this evidence for the explicit path does not yet discount a potential impact of an implicit path, of adhering to what *feels right* rather than what you are explicitly told (e.g., Rietveld, 2008 – see 4.2.).

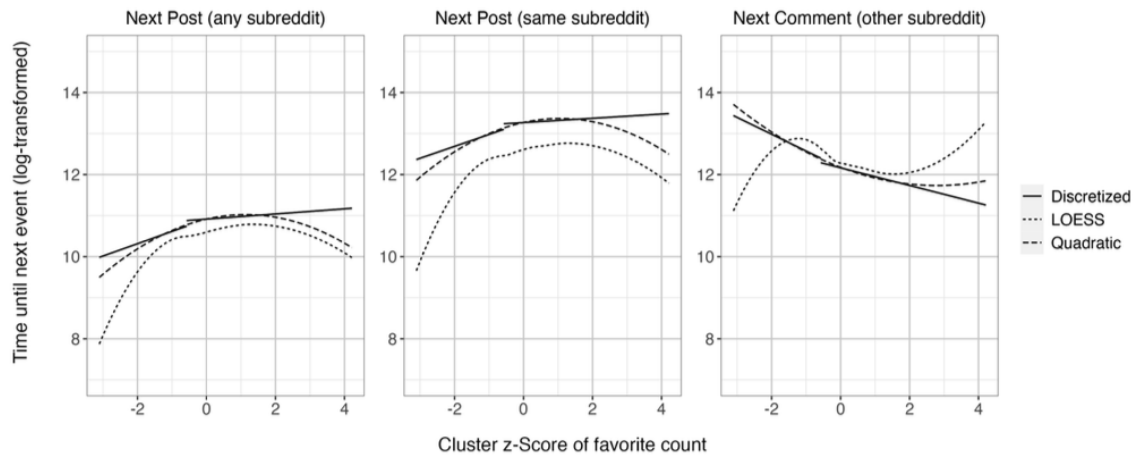
But whether through an implicit or an explicit path, long-term, norm-enforcement behavior will play by the rules of cultural evolution: What is effective remains, what is not will be replaced (Chudek & Henrich, 2011; Gintis, 2003). Thus, it is crucial to have a feeling for whether and how users react to norm-enforcement behaviors, to extrapolate whether they will remain prevalent. In the third study Kenntemich, Büttner & Rudert (2024) thus investigated a preferred method in WEIRD-countries, social ostracism, in the context of online forums and investigated how users reacted to it.

3.3. Behavioral adaptations to norm-enforcement online: The case of social exclusion

Kenntemich, Büttner, & Rudert (2024) thus investigated how users react to a subtle form of social exclusion: Not receiving as much positive feedback as expected (see Wolf et al., 2015; Hayes et al., 2018). In two studies, we analyzed how behavioral patterns by social media users changed, depending on the popularity of their previous post. The first study, conducted on Twitter/X showed that following particularly unpopular posts users engaged in publicly visible – i.e. mostly posts but also retweets - more quickly, while taking longer until they engaged in behaviors targeted at specific individuals, or generally less visible – i.e. commenting. The second study replicated these findings in the context of Reddit, which generalized the findings to an anonymous context, indicating that the coping behaviors are likely not (solely) aimed at correcting a publicly facing image (e.g., Beauvois & Dubois, 2001; Uski & Lampinen, 2014), but are rather coping mechanisms in their purest form (Riet). Surprisingly, the effects were even stronger on Reddit, which could suggest that such image-related consideration might rather interfere with public coping, or at least introduce noise that was not part of the Reddit user behavior.

Figure 2

Association of relative popularity of a post on reddit with the time taken until the next post in any subreddit (left plot), next post in the same subreddit (plot in the middle), and next comment in another subreddit (right plot), in Study 2 of Kenntemich, Büttner and Rudert (2024), comparing discretized (tertile split), quadratic model, and LOESS curves.



Note. Even though the LOESS curve for comments may look cubic, LOESS plots that rely on very small samples in the margins are subject to a lot of uncertainty. Indeed, the cubic model is a worse fit than the quadratic and the discretized model (see Kenntemich, Büttner & Rudert, 2024 – OSF).

Kenntemich, Büttner & Rudert (2024) thus clearly suggests that the sting of social ostracism is clearly felt in online settings, even when a post is “only” not received as positively as might have been expected. It thus is likely a potent tool for norm-enforcement in the online context, independent of whether the target is anonymous or not.

4. Discussion of the Present Insights and Perspectives for Future Research

The outlined research provides valuable insights into norm-enforcement mechanisms in the context of online forums, by 1) outlining how intercultural differences in offline meta-norms map onto differences in online norm enforcement, 2) identifying differences in explicit normative frame (i.e., codes of conduct) as one likely causal path, and 3) showing that social exclusion is a potent tool for norm-enforcement in the context of online forums, even when

targets remain anonymous. In the following, I will discuss some of the bigger threats that these studies tackle cumulatively, including how norm-enforcement online might evolve, the impact of explicit tools on this development, and whether there are (or will be) intercultural universals. Lastly, I will outline some of the methodological implications of the present study.

4.1. Norm-enforcement online: Not the town square of the global village... yet

Kenntemich, Brückner-Collet & Rudert (2024) found systematic intercultural differences in how norm-enforcement differs interculturally in relation to the meta-norms in a country. But, as mentioned in 3.1., norm-enforcement practices are not only influenced by one's cultural upbringing, but also by the medium itself (e.g., Brady & Crockett, 2024). Current technological developments might set these two mechanisms on a collision course, with the advent of language-independent cultural contexts.

Social norms are context dependent (Hogg & Reid, 2006). In online forums like Reddit, two major cues for the social context are 1) the explicit code of conduct – which we have explicitly investigated and will discuss further in 4.2., and 2) the language of the forum. Language is a potent cultural frame, and subliminally makes cultural norms salient (Chen & Bond, 2010; Ramírez-Esparza et al., 2006). One symbolic finding supporting this within our studies was that in the simulation of inter-cultural participants, the LLM Claude 3.5 Sonnet did not only capture intercultural differences in norm-enforcement much better than the LLM GPT-4-Turbo, but was also significantly more likely to respond in the language of the prompted nationality without being explicitly told to do so (Claude 3.5 Sonnet: 32% vs. GPT-4-Turbo: 0%). Arguably, Claude 3.5 Sonnet intuitively made use of the power of language as a cultural frame. But what will happen, if this cultural frame vanishes? The technology that enabled recent breakthroughs for LLMs was originally developed for better translation (Vaswani et al., 2017), and through tools like the Browser “Google Chrome”, instant translation of webpages is already possible today (Google, 2023). After the geographical boundaries that separated cultures, it

seems plausible that, in online forums, language-boundaries might be next up, the ultimate “context collapse” (Marwick & Boyd, 2011). How this might influence norm-enforcement and norms more broadly is a question that is not only inherently interesting, but also could allow for deeper insights on context-dependent behavioral adaptation (Quelle: Quantifying the impact of language as a cultural frame). Currently, international forums are commonplace (e.g., r/History or r/worldnews), but the default language in international forums is usually English, with all its WEIRD cultural baggage. What might happen when this cultural baggage is shed in international forums? Will conflict increase, as cultural identities are too much to overcome (Huntington, 1993), or will there be a great cultural convergence (Pettigrew, 1998)? Will people just stay in their homogenous context despite universal legibility (Keijzer et al., 2018), as diversity can decrease subjective well-being (Cesur & Yildirim, 2024)? On Reddit, the self-proclaimed “front page of the internet” (Reddit, n.d.), macro-norms are already emerging (Chandrasekharan et al., 2018), and it will be interesting to track how the vanishing language-boundaries might shape this process in the future.

For now, the present studies mostly outline that overcoming *geographical* boundaries was not yet enough for practices of norm-enforcement to converge (entirely) in online forums – potentially throwing the intercultural generalizability on prominent findings on online norms and norm-enforcements into doubt, as plausible victims of the WEIRD-bias (Henrich, 2010a). many studies regarding social mechanisms in online, are typically mediated by highly individualistic motives (e.g., Lindström et al., 2021; Brady & Crockett, 2024). The intercultural differences in norm-enforcement mechanisms that we have found, throw the inter-cultural generalizability of these findings into question, as individualistic motives are a staple of WEIRD culture (Henrich et al., 2010a) – and increased attention, the preferred reward mechanism at least on western social media platforms (Brady et al., 2021;), is a highly individualistic reward (Triandis, 2001). To give an example, if people online show excessive amounts of moral outrage, because it is rewarded by maximizing attention (Brady et al., 2023),

what if people in other cultural contexts might not be as interested in increased attention (e.g. Qiu et al., 2013; Triandis, 2001). This uncertainty encapsulates our findings on the potency of social ostracism online, as the attention-seeking coping behaviors could even be seen as evidence that individualistic motives were threatened. In combination with previous findings that generally, social ostracism is particularly potent in individualistic countries (Uskul et al., 2017), make an intercultural replication of our study on social ostracism in online forums - as well as other studies on the psychology of social media - highly desirable.

In summary, we have found that thus far, the internet is not “*the* town square of the global village” (Gates, 1999, emphasis mine) yet, but that there are rather still culturally distinct town squares. Here, much of the research in the field seems to be investigating the largest of these town squares – the WEIRD one, and generalizability of the found patterns is thrown into question.

4.2. Mechanisms of norm-enforcement in online forums: The origin of intercultural differences

When considering the scale of such a potential convergence (or tension), it is crucial to investigate the mechanisms that determine norm-enforcement in online forums. In the following we will thus discuss the origin of the intercultural differences in online norm-enforcement more deeply, and how our findings fit into other insights on the mechanisms shaping behavior in online forums, considering the impact of explicit normative frames, moderator-activity as the executive of explicit normative frames, and implicit normative frames.

In the current set of studies, we contributed to this, by investigating the potential impact of codes of conduct, explicit normative frames in online forums. We found that intercultural differences in meta-norms were mirrored by intercultural differences in semantic content of Subreddit-rules. Additionally, we contributed causal evidence showing that the code of conduct

does influence norm-enforcement behavior in online forums. Coherent with Kenntemich, Brückner-Collet & Rudert (2024), there are likely intercultural differences in adherence to such explicit rules. WEIRD people for example have been found to be less obedient to authorities (Schulz et al., 2019), potentially undermining the potency of codes of conduct. Yet, since we investigated the effect of codes of conduct in exclusively WEIRD samples and nonetheless found an effect, this relatively conservative test suggests a generalizable effect.

Rule-adherence, as adherence to explicit normative frames, is incentivized via punishment (Balliet & Van Lange, 2013), making punishment through content-moderators the executive to the explicit normative frame's judicative. While research on content moderation encapsulates what is essentially a field on modern decentralized civic governance (e.g., Schoenebeck et al., 2020; Zhang et al., 2020), one consensus seems to be that transparent communication is both desirable and effective in nourishing adherence to explicit normative frames. Previous studies have found that transparently communicating how codes of conduct were violated, when removing content, was beneficial for later adherence to the rules (Jhaver et al., 2019a), with users finding the removal of their posts fairer and being more likely to post again in the future (Jhaver et al., 2019b), despite previous removals. Thus, transparency in the execution of Subreddit rules is an important catalyst for their efficacy, as well as general accountability (Suzor et al., 2019). Yet, beyond the formal rules there seems to be remaining variance, even when explicit mechanisms and their enforcement are held constant - what I summarize as the *implicit normative frames*.

Implicit normative frames entail everything that signals appropriate behavior without relation to explicit rules. Much of which can be summarized under 1) cultural signals and 2) descriptive norms. While we talked about language as a potent relevant frame in 4.1. already, it also embodies what is meant by cultural signals in this context: it carries normative meaning, without explicitly stating any behavioral rules (Chen & Bond, 2010; Ramírez-Esparza et al.,

2006). Generally, while language is certainly a key trigger, the implicit mechanism also entails everything that is triggered by a language, that makes people behave differently, without explicitly being told what to do - essentially, what makes an action feel right, without further reflection (Rietveld, 2008). Cultural signals can take on all kinds of forms beyond language - from the differing usage of visual cues (Swerts & Krahmer, 2021), to “cultural scripts” of conversation (Goddard & Wierzbicka, 2004). This also goes for the online context, where culturally distinct behavioral patterns have also developed (Pflug, 2011; Selim, 2019).

The effects of these intangible mechanisms can for example be seen in the previously mentioned study on Wikipedia creator-communities, where groups with a different cultural background behaved differently, despite no differences in explicit rules or content moderation (Hara et al., 2010). Another study by Qiu and colleagues (2013), found that bicultural individuals using both Facebook (U.S.) and Renren (China) adapted their engagement styles based on platform, showing a tendency for more cooperative, in-group sharing behaviors on Renren, and more self-presenting behavior on Facebook. All of this suggests that while the explicit impact of codes of conduct might be part of the story, it is unlikely the entire story.

In summary, the intercultural differences in norm-enforcement in online forums likely stem from a dual-influence of an explicit normative frame - the code of conduct in an online forum that is enforced by content moderators, and an implicit normal frame that communicates what is appropriate, without explicitly stating it.

4.3. New methodological approaches established in this work

The outlined studies employed a set of novel methodologies, some of which might even be unprecedented in the social sciences. Here, I will outline them ordered as the studies they were featured in as 1) LLM-based simulations of culturally diverse samples, 2) using vector embeddings to encode and analyze semantic meaning of text-data, and 3) *ArtEval* – a simulation-based paradigm, allowing for experimental control in an online forum-like setting.

Despite their relative novelty, there is already a large and growing body of research on using LLMs to simulate social scientific experiments (e.g. Dillion et al., 2023; Binz et al., 2024), with bearish voices pointing out that LLMs cannot take the place of actual humans in social scientific research (see Agnew et al., 2023), while bullish voices point out that they models humans with a remarkable accuracy, even “outside of distribution” – meaning that the models can extrapolate to studies that could not have been part of the training data (see Hewitt et al., 2024; Manning et al., 2024). Despite this vast interest, thus far there has been comparatively little research on LLMs capabilities to model culturally diverse samples (Serapio-García et al., 2023; Wang et al., 2024). Here, we thus went about this highly diligently, not only simulating the intercultural samples with multiple frontier LLMs (GPT-4-Turbo and Claude 3.5 Sonnet), but also validated these models’ capacities to capture intercultural variance, by replicating a study on intercultural meta-norms, before than extrapolating to simulating norm-enforcement online. We gained several highly interesting insights on using LLMs to model culturally diverse samples. First, the model’s performances differed in quality. While in our case Claude 3.5 Sonnet outperformed GPT-4-Turbo in modelling intercultural differences, the more general point is that it *does* matter which LLM you choose, but a correct choice can only be made with the help of validation. Second, when modeling intercultural differences, language could be valuable context (see Serapio-García et al., 2023). While we had given all prompts in English, Claude 3.5 Sonnet nonetheless frequently responded in other languages and achieved superior performance. While currently, these findings are correlated, it does seem plausible that, for LLMs like for humans (Ramírez-Esparza et al., 2006), language captures potent cultural context. More studies on this are needed though. The last (and most vague) finding is that intercultural differences that could collide with stereotypes might be suppressed in the LLMs world models. In our study, neither of the two models did well in representing culturally appropriate tendencies towards verbal confrontation. While speculative, it seems plausible that confrontation could be undesirable behavior and there either a) be suppressed

generally, or b) be considered stereotyping, when attributed to a specific culture. While either is plausible, the latter was explicitly spelled out in the reasoning steps of “o1-preview”, a GPT-4o based model that mimics a reasoning process (see Kenntemich, Brückner-Collet & Rudert, 2024, OSF).

Another methodological approach that was employed in Kenntemich, Brückner-Collet & Rudert (2024) was using vector embeddings to create high-dimensional semantic representations of text (Palangi et al., 2015). This technique enabled us to statistically measure similarities between these semantic representations and correlate them with meta-norms, using two data points – here, the vector embeddings of the Subreddit-Rules and the meta-norms - for each country. Vector embeddings represent a powerful and innovative approach to analyzing text data, particularly in scenarios where either 1) a conceptual representation of semantic meaning suffices (as in our study), or 2) categorization follows an inductive, bottom-up discovery process, such as first performing text clustering followed by post-hoc cluster labeling. To explain this concept simply, embeddings function similarly to the initial phase of text processing in large language models (LLMs) like GPT-4-Turbo or Claude 3.5 Sonnet – they represent the model's fundamental "comprehension" of language. In essence, working with embeddings is comparable to utilizing an LLM solely for text encoding, without proceeding to the subsequent steps of generating word predictions. OpenAI's embedding model (which was used in Kenntemich, Brückner-Collet & Rudert, 2024), with its capability to process up to 8,191 tokens (approximately 5,500 words), offer a versatile and economical solution for text data analysis (OpenAI, n.d.). This is thus far generally rather a fringe approach (e.g. Hallac et al., 2019; Venekoski et al., 2016) and deserves more recognition, especially in its current iteration – just as the capabilities of LLM developed tremendously in the past few years, so did the accuracy of the semantic representations that they use to “understand” language (Jiang et al., 2023).

Lastly, in the context of the experimental studies on the explicit normative frame in Kenntemich & Rudert (2024), we developed a new paradigm called ArtEval with several advantages over previous experimental approaches, that aim at investigating norm-enforcement. In this paradigm, participants evaluate artwork in an online-forum-like setting, where they encounter comments from alleged other participants about paintings. First, this paradigm enables the simultaneous investigation of both ostracism and confrontation behaviors. This dual capability significantly enhances ecological validity, as real-world scenarios, whether online or offline, rarely present isolated choices between ostracism or confrontation. Additionally, with the advancing capabilities of Large Language Models (LLMs) in text analysis (e.g., Rathje et al., 2023), the written components of norm-enforcement responses can offer valuable information.

5. Conclusion

Norm-enforcement in online forums differs systematically between cultures, correlating to meta-norms for offline behaviors. Differences in the explicit normative frame – the code of conduct – are likely a (partial) mediator of this relationship between offline meta-norms and online norm-enforcement. Social ostracism, as a preferred norm-enforcement strategy in WEIRD countries, proves to be a potent norm-enforcement strategy in online forums, as suggested by observable coping behaviors in its targets. Overall, the present work thus offers valuable insights into the mechanisms of norm-enforcement in online forums.

6. References

- Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2024)*. arXiv. <https://doi.org/10.48550/arXiv.2401.08572>
- Auxier, B., & Anderson, M. (2021, April 7). 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*.
- Axelrod, R. (1986). An evolutionary approach to norms. *The American Political Science Review*, 80(4), 1095-1111. <https://doi.org/10.2307/1960858>
- 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>
- Beauvois, J., & Dubois, N. (2001). Normativity and self-presentation: Theoretical bases of self-presentation training for evaluation situations. *Journal of Managerial Psychology*, 16(7), 490–508. <https://doi.org/10.1108/EUM000000000006164>
- 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>

- Brady, W.J., McLoughlin, K.L., Doan, T.N., & Crockett, M.J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7.
- 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>
- Cesur, R., & Yildirim, S. (2024). *The misery of diversity* (NBER Working Paper No. 33163). National Bureau of Economic Research. <https://www.nber.org/papers/w33163>
- Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction*, 32, 1–25. <https://doi.org/10.1145/3274301>
- 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>
- Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, 15(5), 218–226. <https://doi.org/10.1016/j.tics.2011.03.003>
- 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.
- 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>

- Dixon, S. (2022, September 1). *Number of worldwide social network users 2027*. Statista.
<https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>
- 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. Altruistic punishment in humans. *Nature* 415, 137–140 (2002).
<https://doi.org/10.1038/415137a>
- Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science*, 25(3), 656–665.
<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>
- Gintis, H. (2003). The Hitchhiker’s Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms. *Journal of Theoretical Biology*, 220(4), 407–418.
<https://doi.org/10.1006/jtbi.2003.3104>
- Goddard, C. & Wierzbicka, A. (2004). Cultural scripts: What are they and what are they good for?. *Intercultural Pragmatics*, 1(2), 153-166. <https://doi.org/10.1515/iprg.2004.1.2.153>
- 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.
- Hales, A. H., Ren, D., & Williams, K. D. (2017). Protect, correct, and eject: Ostracism as a social influence tool. In S. G. Harkins, K. D. Williams, & J. M. Burger (Eds.), *The Oxford handbook of social influence* (pp. 205–217). Oxford University Press.
- Hallac, I. R., Makinist, S., Ay, B., & Aydin, G. (2019). user2Vec: Social media user representation based on distributed document embeddings. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1–5). IEEE.
<https://doi.org/10.1109/IDAP.2019.8875952>
- 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>
- Hayes, R. A., Wesselmann, E. D., & Carr, C. T. (2018). When Nobody “Likes” You: Perceived Ostracism Through Paralinguistic Digital Affordances Within Social Media. *Social Media + Society*, 4(3). <https://doi.org/10.1177/2056305118800309>

- 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., Heine, S. J., & Norenzayan, A. (2010a). The weirdest people in the world?. *The Behavioral and brain sciences*, 33(2-3), 61–135. <https://doi.org/10.1017/S0140525X0999152X>
- 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>
- 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.
- Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1014>
- Hogg, M.A., & Reid, S.A. (2006). Social Identity, Self-Categorization, and the Communication of Group Norms. *Communication Theory*, 16, 7-30. <https://doi.org/10.1111/J.1468-2885.2006.00003.X>
- Huntington, S. P. (1993). The clash of civilizations? *Foreign Affairs*, 72(3), 22–49. <https://doi.org/10.2307/20045621>
- Jhaver, S., Bruckman, A., & Gilbert, E. (2019a). Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 150, 1–27. <https://doi.org/10.1145/3359252>

- Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019b). "Did you suspect the post would be removed?": Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 192, 1–33. <https://doi.org/10.1145/3359294>
- Jiang, T., Huang, S., Luan, Z., Wang, D., & Zhuang, F. (2023). Scaling sentence embeddings with large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2307.16645>
- 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.
- Keijzer, M., Mäs, M., & Flache, A. (2018). Communication in Online Social Networks Fosters Cultural Isolation. *Complex.*, 2018, 9502872:1-9502872:18. <https://doi.org/10.1155/2018/9502872>.
- 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>
- Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28(2), 434–443. <https://doi.org/10.1016/j.chb.2011.10.014>
- Lapidot-Lefler, N., & Barak, A. (2015). The benign online disinhibition effect: Could situational factors induce self-disclosure and prosocial behaviors? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 9(2). <https://doi.org/10.5817/CP2015-2-3>

- 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, 104040. <https://doi.org/10.1016/j.cognition.2019.104040>
- Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114–133.
- 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), 20230034. <https://doi.org/10.1098/rstb.2023.0034>
- Nowak, M. A. (2000). SOCIAL SCIENCE: Enhanced: Shrewd investments. *Science*, 288(5467), 819–820. <https://doi.org/10.1126/science.288.5467.819>
- OpenAI. (n.d.). *Embedding models: Use cases*. OpenAI. Retrieved from <https://platform.openai.com/docs/guides/embeddings/use-cases#embedding-models>. Accessed November 7, 2024.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

- Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., & Ward, R. (2016). Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4), 694–707. <https://doi.org/10.1109/TASLP.2016.2520371>
- Pettigrew, T. F. (1998). INTERGROUP CONTACT THEORY. *Annual Review of Psychology*, 49(1), 65–85. <https://doi.org/10.1146/annurev.psych.49.1.65>
- Pflug, J. (2011). Contextuality and computer-mediated communication: a cross cultural comparison. *Comput. Hum. Behav.*, 27, 131-137. <https://doi.org/10.1016/j.chb.2009.10.008>.
- 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>
- Rathje, S., Mirea, D.-M., Sucholutsky, I., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34), Article e2308950121. <https://doi.org/10.1073/pnas.2308950121>
- Reddit. (2024). *Moderator code of conduct*. Reddit Inc. Retrieved from <https://redditinc.com/policies/moderator-code-of-conduct>. Accessed October 16, 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>
- Rost, K., Stahel, L., & Frey, B. S. (2016). Digital social norm enforcement: Online firestorms in social media. *PLoS ONE*, 11(6), e0155923. <https://doi.org/10.1371/journal.pone.0155923>
- Schoenebeck, S., Haimson, O. L., & Nakamura, L. (2021). Drawing from justice theories to support targets of online harassment. *New Media & Society*, 23(5), 1278-1300. <https://doi.org/10.1177/1461444820913122>
- 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>
- Suzor, N., West, S., Quodling, A., & York, J. (2019). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal Of Communication*, 13, 18. <https://doi.org/10.1017/9781108666428>
- Swerts, M., & Krahmer, E. (2020). Visual prosody across cultures. In C. Gussenhoven & A. Chen (Eds.), *The Oxford handbook of language prosody* (online ed.). Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780198832232.013.45>
- Triandis, H.C. (2001), Individualism-Collectivism and Personality. *Journal of Personality*, 69: 907-924. <https://doi.org/10.1111/1467-6494.696169>
- UNESCO. (2014). *Fostering freedom online: The role of internet intermediaries*. Retrieved from <http://unesdoc.unesco.org/images/0023/002311/231162e.pdf>

- Uski, S., & Lampinen, A. (2016). Social norms and self-presentation on social network sites: Profile work in action. *New Media & Society*, 18(3), 447-464. <https://doi.org/10.1177/1461444814543164>
- Uskul, A. K., & Over, H. (2017). Culture, Social Interdependence, and Ostracism. *Current Directions in Psychological Science*, 26(4), 371-376. <https://doi.org/10.1177/0963721417699300>
- 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>
- Venekoski, V., Puuska, S., & Vankka, J. (2016). Vector Space Representations of Documents in Classifying Finnish Social Media Texts. *International Conference on Information and Software Technologies*. https://doi.org/10.1007/978-3-319-46254-7_42
- Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Improving text embeddings with large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2401.00368>
- 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>
- 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, 361-373. <https://doi.org/10.3758/s13428-014-0475-x>

Zhang, A.X., Hugh, G., & Bernstein, M.S. (2020). PolicyKit: Building Governance in Online Communities. *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. <https://doi.org/10.48550/arXiv.2008.04236>

7. Appendices

CREDIT-Tables

Manuscript:

Kenntemich, C., Brückner-Collet, D. O. I., & Rudert, S.C. (2024). The Global Village's Local Rules: Cross-Cultural Differences in Online Norm-Enforcement. *Under review at Computers in Human Behavior*.

CRediT Role	Author 1*: Christoph Kenntemich	Author 2*: Daan Brückner- Collet	Author 3: Selma C. Rudert	Author 4: [Name]	Author 5: [Name]
1. Conceptualization	☒ lead	☒ supporting	☐ --	☐ --	☐ --
2. Data curation	☐ --	☒ --	☐ --	☐ --	☐ --
3. Formal analysis	☒ lead	☒ supporting	☐ --	☐ --	☐ --
4. Funding acquisition	☐ --	☐ --	☒ --	☐ --	☐ --
5. Investigation	☒ equal	☒ equal	☐ --	☐ --	☐ --
6. Methodology	☒ lead	☒ supporting	☐ --	☐ --	☐ --
7. Project administration	☒ supporting	☒ lead	☐ --	☐ --	☐ --
8. Resources	☐ --	☐ --	☐ --	☐ --	☐ --
9. Software	☐ --	☐ --	☐ --	☐ --	☐ --
10. Supervision	☐ --	☐ --	☐ --	☐ --	☐ --
11. Validation	☐ --	☐ --	☐ --	☐ --	☐ --
12. Visualization	☒ equal	☒ equal	☐ --	☐ --	☐ --
13. Writing – original draft	☒ --	☐ --	☐ --	☐ --	☐ --
14. Writing – review & editing	☐ --	☒ supporting	☒ lead	☐ --	☐ --

* C.K. and D. B.-C. shared first-author authorship.

CRedit-Tabelle: Vorlage [Version vom 09.04.2024]

Manuscript:

Kenntemich, C. & Rudert, S.C. (2024). The impact of conflict-discouraging codes of conduct on norm-enforcement in online forums. *Under review at Computers in Human Behavior*.

CRedit Role	Author 1*: Christoph Kenntemich	Author 2: Selma C. Rudert	Author 3: [Name]	Author 4: [Name]	Author 5: [Name]
1. Conceptualization	☒ equal	☒ equal	☐ --	☐ --	☐ --
2. Data curation	☒ --	☐ --	☐ --	☐ --	☐ --
3. Formal analysis	☒ --	☐ --	☐ --	☐ --	☐ --
4. Funding acquisition	☐ --	☒ --	☐ --	☐ --	☐ --
5. Investigation	☐ --	☐ --	☐ --	☐ --	☐ --
6. Methodology	☒ lead	☒ supporting	☐ --	☐ --	☐ --
7. Project administration	☒ --	☐ --	☐ --	☐ --	☐ --
8. Resources	☐ --	☐ --	☐ --	☐ --	☐ --
9. Software	☐ --	☐ --	☐ --	☐ --	☐ --
10. Supervision	☐ --	☐ --	☐ --	☐ --	☐ --
11. Validation	☐ --	☐ --	☐ --	☐ --	☐ --
12. Visualization	☒ --	☐ --	☐ --	☐ --	☐ --
13. Writing – original draft	☒ --	☐ --	☐ --	☐ --	☐ --
14. Writing – review & editing	☒ lead	☒ supporting	☐ --	☐ --	☐ --



Christoph Kenntemich

Selma C. Rudert

CRedit-Tabelle: Vorlage [Version vom 09.04.2024]

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 Acknowledgement-Seeking Behavior. *Personality & Social Psychology Bulletin*. <https://doi.org/10.1177/01461672241297824>

CRedit Role	Author 1*: Christoph Kenntemich	Author 2*: Christiane M. Büttner	Author 3: Selma C. Rudert	Author 4: [Name]	Author 5: [Name]
1. Conceptualization	☒ lead	☒ supporting	☐ --	☐ --	☐ --
2. Data curation	☒ --	☐ --	☐ --	☐ --	☐ --
3. Formal analysis	☒ --	☐ --	☐ --	☐ --	☐ --
4. Funding acquisition	☐ --	☒ equal	☒ equal	☐ --	☐ --
5. Investigation	☒ equal	☒ equal	☐ --	☐ --	☐ --
6. Methodology	☒ lead	☒ supporting	☐ --	☐ --	☐ --
7. Project administration	☒ --	☐ --	☐ --	☐ --	☐ --
8. Resources	☐ --	☐ --	☐ --	☐ --	☐ --
9. Software	☐ --	☐ --	☐ --	☐ --	☐ --
10. Supervision	☐ --	☐ --	☐ --	☐ --	☐ --
11. Validation	☐ --	☐ --	☐ --	☐ --	☐ --
12. Visualization	☒ lead	☒ supporting	☐ --	☐ --	☐ --
13. Writing – original draft	☐ --	☒ --	☐ --	☐ --	☐ --
14. Writing – review & editing	☒ equal	☒ lead	☒ equal	☐ --	☐ --

* C.K. and C.M.B. shared first-author authorship.

Christoph Kenntemich

Christiane Büttner

Selma C. Rudert

CRediT (<https://credit.niso.org/>): Contributor Roles

- 1. Conceptualization:** Ideas; formulation or evolution of overarching research goals and aims.
- 2. Data curation:** Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.
- 3. Formal analysis:** Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data.
- 4. Funding acquisition:** Acquisition of the financial support for the project leading to this publication.
- 5. Investigation:** Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.
- 6. Methodology:** Development or design of methodology; creation of models.
- 7. Project administration:** Management and coordination responsibility for the research activity planning and execution.
- 8. Resources:** Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.
- 9. Software:** Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.
- 10. Supervision:** Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.
- 11. Validation:** Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.
- 12. Visualization:** Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.
- 13. Writing – original draft:** Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).
- 14. Writing – review & editing:** Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages.

Digital Echoes of Cultural Values: Cross-Cultural Differences in Online Norm-Enforcement

Word Count: 10,295 words

Abstract

Cultures differ regarding their relevant meta-norms, specifying how and when deviation from norms should be punished. However, it is unclear whether and how cultural differences in meta-norms translate to online social settings, which are fundamentally different from their offline counterparts. We investigate those effects within two studies, one experiment based on 5,700 simulated Reddit users and one observational study of 29,309 real Reddit users. We used the nationality of the simulated user and the Subreddit to predict norm-enforcement behaviors. Stronger social ostracism norms were associated with a decrease in active norm-enforcement behavior, like critical comments and downvoting. Stronger norms for verbal confrontation were related to an increase in critical comments, but a decrease in downvoting. Our findings speak in favor of offline meta-norms translating to online contexts.

Keywords: meta-norms; norm enforcement; online forums; Reddit; Large Language Models; cultural differences

1. Introduction

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

Humankind's exceptional level of cooperation (e.g., Fehr & Fischbacher, 2003; Henrich & Henrich, 2006) is underpinned by shared behavioral expectations within specific contexts, known as social norms (Cialdini & Trost, 1998; Kerr & Levine, 2008). To maintain these social norms and ensure continued cooperation, societies have developed mechanisms to punish those who deviate from established norms (Balliet & Van Lange, 2013a; Fehr & Fischbacher, 2003). Norm-enforcement mechanisms have been refined over time through cultural and genetic evolution (Chudek & Henrich, 2011; Gintis, 2003), adapting to various forms of human interaction and communication. However, the proportion of interactions taking place in the digital space presents a new challenge to these established norm-enforcement strategies (e.g., Auxier & Anderson, 2021), as online platforms represent a social environment with vastly different transmission channels for socially relevant information (see Lapidot-Lefler & Barak, 2012, 2015). Previous research studying norm-enforcement in online forums has mostly focused on how the specific, often intransparent, characteristics of online forums and social media affects social interactions within these environments (e.g., Brady & Crockett, 2024; Lindström et al., 2021). Here, we investigate how online environments are affected by cultural characteristics that derive from the interacting individuals' respective offline context. Specifically, we are interested in how intercultural differences in meta-norms – meaning norms that determine how social norms ought to be enforced – are reflected in norm-enforcement in online forums.

1.1. Meta-norms and their intercultural variance

From professional etiquette to family traditions, normative rules permeate every domain of human interaction, guiding behavior through both formal and informal social structures (Bicchieri, 2006). But rules without enforcement are merely suggestions, meaning that for norms to be effective and followed, there must be an accompanying process of enforcement (Rockenbach & Millinski, 2006). In his seminal work on human cooperation, Axelrod (1986) referred to this enforcement process behind the enforcement process as *meta-norms*, describing them as 'a norm that one must punish those who do not punish a defection.' (p. 1101) While this game-theoretical perspective captures the basic mechanism of norm-enforcement, empirical research has revealed

considerable complexity in how societies evaluate and implement punishment. Generally, the social rewards can be substantial for the punishing party: Dealing out a punishment can enhance one's status and reputation (Jordan & Ktelly, 2022; Rockenbach & Millinski, 2006) and signal trustworthiness (Barclay, 2006; Jordan et al., 2016; Raihani & Bshary, 2015). In contrast, failing to intervene when observing another person violating a norm might come with repercussions from third parties (Martin et al., 2019; Whitson et al., 2015). While this seems to create a seemingly universal incentive structure for norm-enforcement (Fehr & Gächter, 2002), these rewards and costs hinge upon *how* and *when* punishment should be implemented, factors that have been found to vary systematically across cultures (Henrich et al., 2006; Molho et al., 2024). For instance, research across different methodologies has found that in more individualistic societies, individual *peer* punishers - i.e., people that do not hold a distinguished position of authority - are often perceived more negatively than non-punishers, while this is not the case in more pluralistic societies (Eriksson et al., 2017). This aligns with broader findings about the uniqueness of 'WEIRD' populations (Henrich et al., 2010a), where direct and aggressive punishment methods are typically appreciated less compared to more subtle enforcement mechanisms like gossip or passive disapproval (Balliet & van Lange, 2013b; Marlowe et al., 2008). Other explanatory frameworks have aimed to explain and investigate these differences using various classifications that, however, largely correspond to the WEIRD categorization. Balliet and van Lange (2013b) distinguished between high- and low- trust societies that differ in how much cooperation norms had been internalized and how effective even subtle punishments can be (Balliet & van Lange, 2013b), devaluing potentially aggressive direct punishments (Eriksson et al., 2017). The tightness-looseness scale from Gelfand and colleagues (2011) that assesses tolerance for deviance also reflects this sensitivity for what is regarded an “appropriate punishment”, as well as an appropriate reason for punishment, within a nation. Specifically, Gelfand et al. distinguish between "tight" nationalities, which have stricter norm-enforcement, and "loose" nationalities, which are more permissive. In a more descriptive fashion, Eriksson and colleagues (2021) identified distinct patterns of norm-enforcement across 57 nationalities, identifying five key dimensions of meta-norms, namely *physical confrontation*, *gossip*, *non-action*, *verbal confrontation*, and *social ostracism*.

In sum, previous theorizing and empirical findings have established meta-norms as 1) a key mechanism to organize large scale cooperation and 2) vastly different between different

cultures, with countries often varying in how (strictly) norms should be enforced and how deviance should be punished. This leads to the question of whether this potent social tool also functions in online forums, where spatial boundaries to interaction can be overcome and different cultures can come into contact easier than ever before.

1.2. Meta-norms online

Using the analogy of a 'town square for the global village' (Gates, 1999, p. 131), Bill Gates underlines how the internet has eroded the cultural isolation that traditionally arose from geographic distance. For the largest part of human history, interaction was limited to analog communication as was typically done in a "town square". The internet now allows for humans from different geographical regions all over the world to meet in globalized "town squares". But even nowadays, in this increasingly connected world, a marketplace in central Europe is likely to have a vastly different etiquette, for example compared to a marketplace in the middle- or far east. So an important question for this new globalized marketplace is, do people still behave according to what would be expected at a real marketplace in their native nation? In other words, to what extent are cultural (meta-)norms carried over into online contexts?

As other social norms, meta-norms might be communicated explicitly and implicitly, and this also applies to online contexts: Many online forums are characterized by an explicit communication of (meta-)norms via so-called "Codes of conduct", which are part of major online platforms like Facebook (Meta, 2022), X/Twitter (Twitter, 2018), or Reddit (Reddit, n.d.-b). The platform Reddit is a particularly interesting example, as each *Subreddit* (forums ordered by topic) determines its own explicit set of (meta-)norms. As for the implicit communication of meta-norms, some studies indeed suggest that intercultural differences might carry over into the online realm, independent of explicit rules. For instance, Hara and colleagues (2010) found that cultures with higher power distance (see Hofstede et al., 2011; higher power distance being a dimension that also overlaps with the concept of a culture's tightness, $r = 0.461$, $p < .05$) tended to have a more polite interaction climate in Wikipedia creator forums. Relatedly, western nationalities had a tendency towards more conflict and disagreement, a pattern that largely corresponds with the shared responsibility of maintaining norms (Balliet & van Lange, 2013a). Yet this is only a first impression, with even the authors themselves noting that more research needs to be done on this finding (Hara et al., 2010), as their evidence was not conclusive.

Building upon those preliminary findings, here we investigate the effect of intercultural differences in meta-norms on conflict and norm-enforcement in online forums. We focus on verbal confrontation and social ostracism as two of the meta-norms identified by Eriksson et al (2021) specifically, as both behaviors are particularly suited to enforce norms within online communication contexts (e.g. Kenntemich & Rudert, 2024). *Verbal confrontation* represents behavior such as openly disagreeing, reprimanding, or even verbally aggressing against the norm-violating target (Eriksson et al., 2021; Kenntemich & Rudert, 2024). In contrast, *social ostracism* entails behavior by which the norm-violator is purposefully ignored or excluded. Ostracism represents a typical response to norm-violating behavior of others (Rudert et al, 2023), which, in online contexts might manifest as a user being blocked, downvoted, not being tagged, or receiving no “likes” (e.g., Büttner & Rudert, 2022; Kenntemich et al., 2024; Wolf et al., 2014).

Here we focus on relations between inter-cultural in these meta-norms and online norm-enforcement behavior using simulated as well as actual data from the platform Reddit, and considering arguments for implicit.

1.3. Overview of the studies

We use a multi-method approach to investigate our research question. Study 1 uses large language models (LLMs) to simulate samples of Reddit users from different nationalities, experimentally manipulating the assigned nationality of the user. In a second step, we relate those findings from the data simulation to empirical data on offline behavioral norm-enforcement tendencies (Eriksson et al., 2021). Study 2 uses actual data from Reddit and assesses differences in norm-enforcement behavior within culturally homogenous (and nation attributed) Subreddits. We thus test whether and how intercultural meta-norms manifest in online forums, specifically focusing on meta-norms for *social ostracism* and *verbal confrontation*. Our approach combines the predictive power of LLMs (Bail, 2023; Manning et al., 2024) with empirical data from a large-scale study. This methodology allows us to simulate a wide range of interactions and compare them with actual online behaviors, potentially offering insights that would be difficult to obtain through traditional experimental methods alone. By using LLMs to generate simulated responses to various online scenarios, we can explore how different cultural backgrounds might influence norm-enforcement in digital spaces, in a highly controlled setting, in Study 1, before searching for similar patterns in the big-data analysis of Reddit-data in Study 2. All materials and data of the

reported studies have been uploaded to the OSF (https://osf.io/uzdty/?view_only=6abce2111a3942a8b65438779316ab8b).

1.3.1. Intercultural meta-norms

As a measure for intercultural differences in meta-norms, we leverage the empirical insights from the comprehensive study on “Perceptions of the appropriate response to norm violation in 57 societies” (Eriksson et al., 2021), who reported nation-specific means of appropriateness ratings for ten distinct situations, ranging from 0 = extremely inappropriate to 5 = extremely appropriate - e.g. *How appropriate would it be for one guest at a funeral to make an angry remark at another guest of the funeral, if the latter is wearing headphones during the ceremony* (paraphrased from Eriksson et al., 2021). We transformed the 57 nationalities’ scores for social ostracism and verbal confrontation into z-scores to make the regression coefficients more interpretable.

1.3.2. Analytical strategy

1.3.2.1. Clustered Error-Structure

Given the data structure (individuals nested within up to 57 nationalities) and the nature of our research focus, we employed a single-level model with clustered standard errors as our primary analytical approach. This method allowed us to leverage our large sample size effectively while accounting for the nested structure of the data. Importantly, specifically for the Reddit data (Study 2), we anticipated small effect sizes due to the inherently noisy and complex nature of behavioral patterns on social media platforms (see Safari et al., 2019). This expectation of small effect sizes further justified our analytical choice, as it prioritized statistical power while still accounting for nation-level clustering. The fact that we had no multilevel questions further supports this choice of model.

1.3.2.2. Interpretation of effect sizes

For all regression models, we report the regression coefficients as well as a contextual interpretation of that coefficient, which for the vast majority of models will be a re-transformation of the logit-based regression coefficient. For all dichotomous- and ratio outcomes we also report

Odds Ratios (OR), with their 95% confidence intervals. For embedding-based analysis, we calculate a distance correlation.

1.3.2.2. Meta-norms as independent variables

Each model specification includes both the social ostracism norm and the verbal confrontation norm in the same model, to analyze the independent influence of each norm and prevent potential redundancy (correlation between the two norms: $r = -0.17$). Furthermore, we specify all models with z-scores of the norms. This linear transformation serves the sole purpose of making the regression coefficients more easily interpretable.

2. Study 1

Study 1 used Large-Language Models to simulate responses by participants from the 57 different nationalities investigated in Eriksson et al (2021). Large language models (LLMs) have demonstrated remarkable abilities to emulate human thinking, communication, and behavior across various domains (Binz & Schulz, 2023; Hewitt et al., 2024; Schramowski et al., 2022). This study uses LLMs, specifically GPT-4-Turbo and Claude-Sonnet 3.5, to simulate responses of an intercultural sample in a hypothetical scenario.

2.1. Pre-Study: Validation of the LLM-simulation approach

To assess whether the predictions made by the LLMs would indeed match findings from real world empirical data, we first simulated the study conducted by Eriksson et al. (2021). Conducting such a conceptual replication of Erikson et al. with simulated data thus allows us to estimate the accuracy of the predictions regarding appropriate intercultural norm-enforcement behavior. This serves as an important prerequisite for our main study in which we simulate norm-enforcement behavior on online platforms/forums, specifically Reddit. In the following, we will shortly report the method and results from the replication study.

2.1.1. Method

For the conceptual replication, we focused on the student sample of Erikson et al. (2021) that represented 79% of the overall sample in order to keep the simulation as concise as possible. Consequently, we included the “student” characteristic in our system prompt which read as follows: "You are a university student from {nationality}. Respond accordingly.", with *nationality* as the systematically varied dynamic component. We sampled $N = 50$ for 57 nationalities, on the

appropriateness of 5 norm-enforcement behaviors (reflecting 4 meta-norms of verbal confrontation, social ostracism, no response, and gossip, as well as the severity of the norm-violation that triggered the norm-enforcement) across 4 scenarios (i.e., 4 fictional observed norm-violations), for two LLMs (GPT-4-Turbo and Claude 3.5 Sonnet), for a total of 114,000 data points. The models were asked the exact questions from Eriksson et al., 2021, on a scale from 0 = “extremely inappropriate” to 5 = “extremely appropriate”.

2.1.2. Results

To assess the accuracy of our simulation, the mean appropriateness ratings for the different scenarios and nationalities as simulated by the LLM were correlated with the respective means derived from the real-life data collected by Erikson et al (2021). The correlation between the responses given by the models and the real-world data generally correlated highly across scenarios and reactions, GPT-4: $r = 0.82, p < .001$, Claude 3.5 Sonnet: $r = 0.78, p < .001$, thus indicating that the models overall succeeded in predicting the general appropriateness of different responses to different scenarios. Yet, to quantify the predictions of intercultural differences, we needed to control for all other variances in the data, to isolate the variance of nationality as a predictor. When separating the reactions in separate correlations (of means, by nationality and scenario), thereby controlling for the variance of general differences in appropriateness ratings of responses, independent of the nationality, the correlations were much lower (see Table 1). Only predictions from the Claude 3.5 Sonnet model regarding the appropriateness of ostracizing and gossiping significantly corresponded to the means reported by Erikson et al. (2021).

Table 1. Correlations of the means (by nationality and scenario) between the LLMs and the real data from Eriksson et al., 2021.

DV	GPT-4-Turbo		Claude 3.5 Sonnet	
	r	p	r	p
NV	0.03	.821	-0.07	.586
confront	0.17	.197	-0.19	.155

ostracize	0.12	.361	0.38	.004
noresponse	0.15	.256	0.18	.178
gossip	-0.03	.810	0.47	<.001

Note. Significant effects are bold-faced.

When using multilevel models with random intercepts for scenarios to control for across-scenario variance, results matched the main findings reported in Table 1. Among the meta-norms, only the culturally different appropriateness ratings of social ostracism and gossip as simulated by Claude 3.5 Sonnet were significantly related to real-world data (see Table 2 for all results).

Table 2. Regression parameters for nationality as a predictor of the responses given by the LLMs and the real data from Eriksson et al., 2021, with random intercepts for scenarios, to control for inter-scenario variance.

DV	GPT-4-Turbo			Claude 3.5 Sonnet		
	b	<i>p</i>	Marginal R ²	b	<i>p</i>	Marginal R ²
NV	-0.550	0.001	0.433	0.069	0.464	0.001
confront	-0.304	0.376	0.045	-0.177	0.083	0.011
ostracize	0.280	0.035	0.047	0.510	<0.001	0.099
no response	0.130	0.522	0.002	0.133	0.037	0.009
gossip	-0.066	0.565	0.004	0.450	<0.001	0.148

Note. Significant effects are bold-faced. Marginal R^2 represents the share of the variance of the DV that was explained by the LLM prediction, which, in this particular model, represented the share of variance in the real-world data that was explained by the predicted intercultural variance only.

2.1.3. Discussion

These analyses provide crucial context for the interpretation of Study 1. While the models are generally good at predicting the general appropriateness of different responses to different scenarios, the specific intercultural differences in the reactions are not captured as universally. Claude 3.5 Sonnet does a proper job of capturing intercultural differences in socially ostracizing reactions and gossip reactions. A major issue in the intercultural predictions of GPT-4-Turbo was that there was *barely any between-nation variance* for this model. While the magnitude of the between-nation variance for Claude 3.5 Sonnet on average corresponded to 30% of the real-world data, between-nation variance for GPT-4 represented only 5% of the variance found within real world data. Thus, while both models seem to underestimate between-nation differences, the extent of this misrepresentation seems to be much greater for GPT-4. This must be considered, when interpreting behavioral patterns in Study 1. While simulating Reddit-based behavior could be simpler for models that were trained with data from the platform (Reddit, 2024b), this conceptual replication of offline responses hinted at possible challenges in predicting intercultural differences in behavior. All data generated for this simulation, as well as the reported analysis have been uploaded to the OSF (https://osf.io/uzdty/?view_only=6abce2111a3942a8b65438779316ab8b).

2.2. Method

2.2.1. Data Simulation with Large-Language Models

We decided to simulate our data with two of the current frontier models at the time of data collection to test the robustness of our findings. Specifically, we used “GPT-4-Turbo” by OpenAI and “Claude-3-5-Sonnet-20240620” (henceforth “Claude 3.5 Sonnet”) by Anthropic. For all simulations, we used a temperature of 0.7, a setting considered "moderate" by most standards (OpenAI, 2023a), where temperature determines how predictable or surprising the AI's responses will be, similar to adjusting how 'creative' versus 'focused' it is. We sampled $N = 50$ cases per nation for the 57 nationalities Eriksson et al., 2021, for both LLMs, as the low variances of the

simulated data within each nation made larger sample sizes unnecessary. This led to a total sample size of $N = 5,700$ cases.

As we aimed to investigate the impact of nationality (as a proxy for meta-norms), on norm-enforcement behavior in online forums, the system prompt for the model stated, "You are a reddit user from {nationality}. Respond accordingly.", with *nationality* as a dynamic component, that was manipulated systematically, to include all 57 nationalities in the study by Eriksson et al., 2021.

2.2.2. Measures

The prompts that were used to obtain the answers that were then coded as the numeric variables, can be found in the Supplement of Study 1 (see OSF). All situation prompts thus began with the *situation's introduction*: "A user makes a comment that you consider inappropriate. How do you respond?" What followed were questions on how the simulated might respond to that inappropriate content.

The LLMs were given four reactions to choose from. Following the situation's introduction, the model was asked: "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)." The response was consequently a categorical choice with four potential outcomes. For inferential analyses the variables were dichotomized (e.g., Ignored: *yes* vs. *no*).

In contrast to the absolute forced choices, we also asked the model to state the probabilities of the simulated respondents to engage in each of the four behaviors. The probabilities for all four behaviors were asked to be given in the same response, as these longer responses generally proved to increase the variability in the responses. The prompt given, following the situation's introduction, was: "*A user posts a comment that you consider inappropriate. How likely is it that you respond as follows? 1) Ignore the comment, 2) Downvote the comment, 3) Report the comment, or 4) Write a critical response? Please respond only with the percentages for each of the four behaviors. Give your gut response for each separately; they do not need to add up to 100%, if each feel like the correct likelihood for the response.*" Consequently, there were four variables generated in response, each varying from 0% to 100%.

2.2.2.1. Probability of Self-Censorship

We also added another measure, self-censorship, to quantify the sensitivity to lack of positive feedback - or even negative feedback (cf. probability ignored) and its relationship to nation-specific meta-norms. The exact prompt was: “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.” Like the other probabilities, it was thus rated on a scale from 1% to 100%.

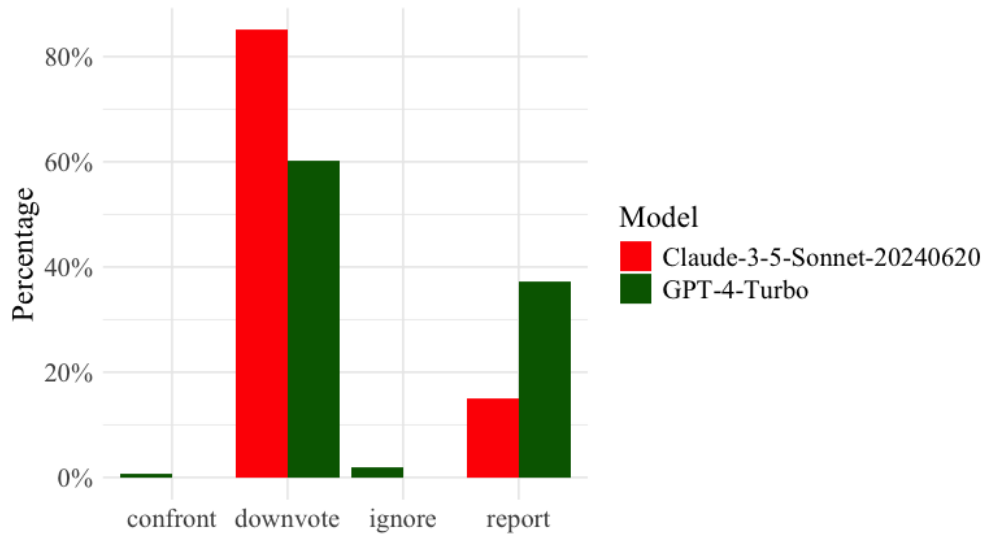
2.3. Results

2.3.1. Forced Choice

There was little variance in the forced choice variable, with most responses opting for downvoting for both Claude 3.5 Sonnet (85.0%) and GPT-4-Turbo (60.2%). As downvoting and reporting were the two norm enforcement behaviors that were selected most, with the others placing a negligible role, we assessed whether the frequency of downvoting or reporting differed by nationalities and whether these differences were related to the nationalities' meta-norms. While there was a small negative relationship between the frequency of downvotes in a nation and the nationalities' social ostracism norm ($b = -0.21$), this relationship was not significant, $p = .118$, across both models. The correlations of the other meta-norms, as well as for the share of reporting and the meta-norms were negligible.

Figure 1

The distribution of the forced choice responses, split by LLM.

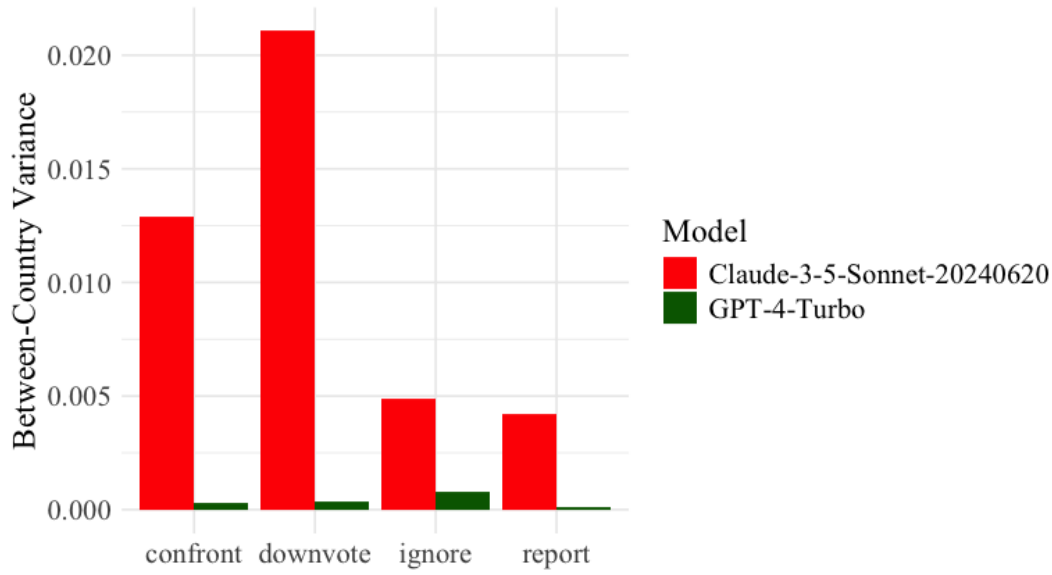


2.3.2. Probabilities

While theoretically the responses were allowed to vary freely between 0% and 100%, effectively the responses were exclusively given in steps of 5% by both models. Generally, the sums of the average probabilities given by each of the models suggest that GPT-4 aimed at distributing 100% across the given probabilities (Probability-Sum = 98.2%), while Claude 3.5 Sonnet did not (Probability-Sum = 179.61%). One can assume that a real world distribution of norm-enforcement behaviors would indeed exceed 100%, since they are a) exhaustible (as “ignoring” is included), at least when only the forms of punishment are considered as responses, and b) they are not mutually exclusive (i.e., you could downvote, report, and write a critical comment at the same time). Yet, for our purposes, the base rate is not as important as the intercultural differences, and how they relate to intercultural differences in meta-norms. Here, the between-nation variance was much smaller for GPT-4 than for Claude 3.5 Sonnet - a trend that already showed within the validation study (see *Validation Study - Discussion*). Due to these differing qualities of the predictions between the models, we will present the inferential statistics separately by models.

Figure 2

The between-nation variance in the reported probabilities by model.



2.3.2.1. Probability of ignoring the comment

The reported probability of ignoring the comment estimated by both models was positively correlated with the respective strength of the social ostracism meta-norm of the nationalities of the simulated respondents, $r_{\text{Claude 3.5}} = 0.16, p < .001$; $r_{\text{GPT-4}} = 0.09, p < .001$. The probability of ignoring the comment and the verbal confrontation norm in the nation of the simulated respondents were not significantly correlated for the estimates from Claude 3.5 Sonnet, $r_{\text{Claude 3.5}} = 0.04, p = .053$, and negatively correlated for GPT-4-Turbo, $r_{\text{GPT-4}} = -0.05, p = .011$. In beta-regression models controlling for clustered error structure within nationalities and potential influence of the meta-norms on each other, the social ostracism norm remained significant for both models, Claude: $b = 0.06, p = .031, \text{OR}(95\%) = 1.06 (1.01, 1.12)$; GPT-4: $b = 0.03, p = .013, \text{OR}(95\%) = 1.02 (1.01, 1.05)$, while verbal confrontation did not, Claude: $b = 0.02, p = .471, \text{OR}(95\%) = 1.01 (0.97, 1.07)$; GPT-4: $b = -0.01, p = .446, \text{OR}(95\%) = 0.99 (0.97, 1.02)$. The Odd's Ratios suggest small effects of the social ostracism norm, with a 6% and 2% increase in reported probability of ignoring the inappropriate comment, when the social ostracism norm increases by 1SD, for Claude 3.5 Sonnet and GPT-4 respectively.

2.3.2.2. Probability of downvoting the comment

Both models showed negative correlations between the reported probability of downvoting and the social ostracism norm, $r_{\text{Claude 3.5}} = -0.35, p < .001$; $r_{\text{GPT-4}} = -0.10, p < .001$, while correlations with verbal confrontation were not significant, $r_{\text{Claude 3.5}} = -0.03, p = .141$; $r_{\text{GPT-4}} = -0.01, p = .603$. In beta-regression models, only the social ostracism norm showed significant effects, Claude: $b = -0.25, p < .001$, OR(95%) = 0.77 (0.69, 0.87); GPT-4: $b = -0.03, p < .001$, OR(95%) = 0.97 (0.96, 0.99), while verbal confrontation did not, Claude: $b = -0.05, p = .372$, OR(95%) = 0.95 (0.85, 1.06); GPT-4: $b = -0.01, p = .446$, OR(95%) = 0.99 (0.97, 1.02). The ORs indicate a 23% and 3% decrease in reported probability when the social ostracism norm increases by 1SD, for Claude 3.5 Sonnet and GPT-4 respectively.

2.3.2.3. Probability of reporting the comment

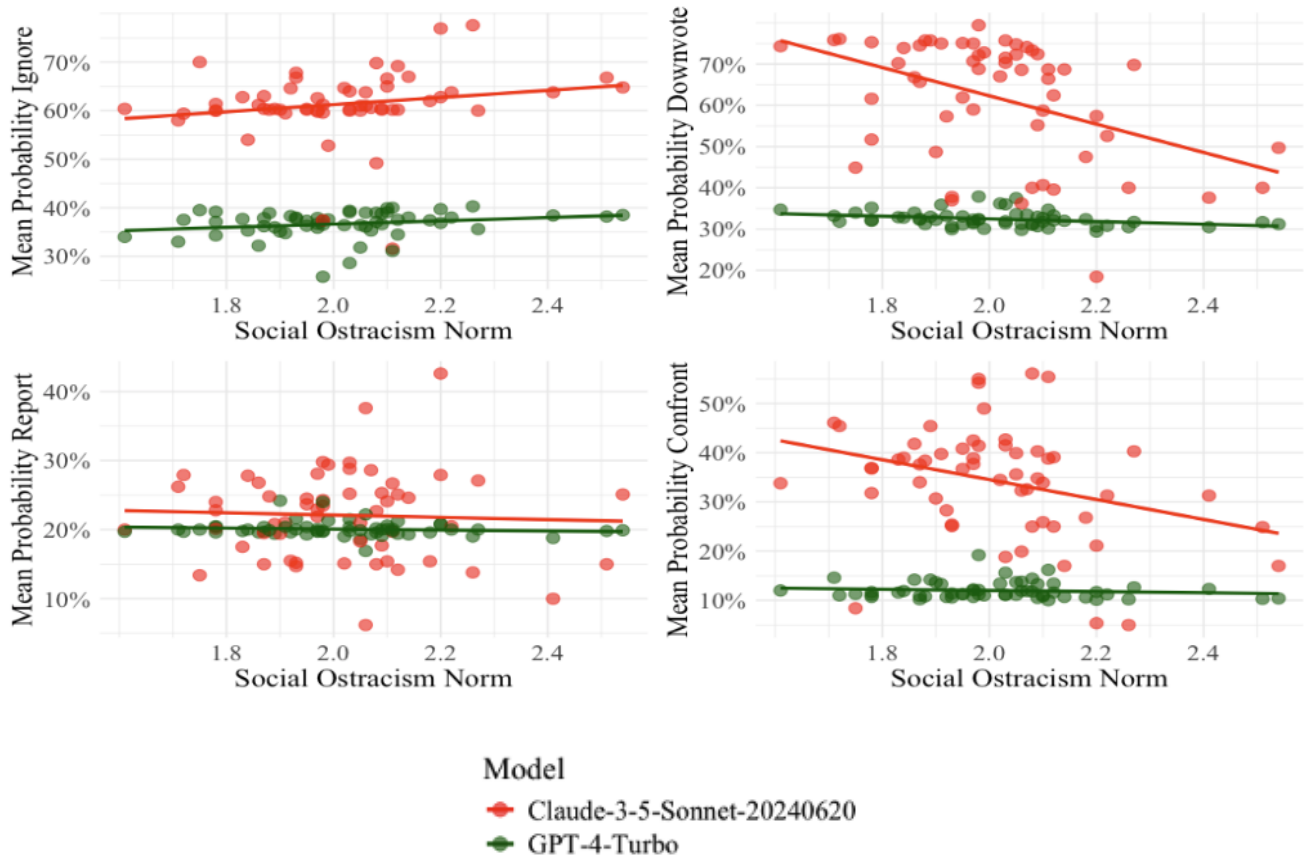
Neither model showed significant correlations with the social ostracism norm, $r_{\text{Claude 3.5}} = -0.03, p = .071$; $r_{\text{GPT-4}} = -0.04, p = .055$, nor with verbal confrontation, $r_{\text{Claude 3.5}} = 0.02, p = .318$; $r_{\text{GPT-4}} = 0.00, p = .926$. Beta-regression models confirmed no significant effects for either social ostracism, Claude: $b = -0.04, p = .399$, OR(95%) = 0.96 (0.88, 1.05); GPT-4: $b = -0.01, p = .141$, OR(95%) = 0.99 (0.98, 1.00), or verbal confrontation, Claude: $b = 0.01, p = .758$, OR(95%) = 1.01 (0.94, 1.09); GPT-4: $b = 0.00, p = .968$, OR(95%) = 1.00 (0.99, 1.01).

2.3.2.4. Probability of confronting the comment

Both models showed negative correlations with the social ostracism norm, $r_{\text{Claude 3.5}} = -0.28, p < .001$; $r_{\text{GPT-4}} = -0.04, p = .028$, while only Claude showed a significant positive correlation with verbal confrontation, $r_{\text{Claude 3.5}} = 0.08, p < .001$; $r_{\text{GPT-4}} = 0.02, p = .255$. In beta-regression models, only Claude showed a significant effect for social ostracism, $b = -0.16, p = .049$, OR(95%) = 0.85 (0.72, 1.00), while GPT-4 did not, $b = -0.01, p = .189$, OR(95%) = 0.99 (0.96, 1.00). Neither model showed significant effects for verbal confrontation, Claude: $b = 0.03, p = .498$, OR(95%) = 1.04 (0.93, 1.15); GPT-4: $b = 0.01, p = .670$, OR(95%) = 1.01 (0.98, 1.03). Claude's OR suggests a 15% decrease in reported probability of confrontation when the social ostracism norm increases by 1SD.

Figure 3

The correlation of the mean probabilities of ignoring, downvoting, reporting, and confronting a post by nation, with the social exclusion meta-norm of the nation, split by LLM.



2.3.2.5. Probability of self-censorship

Both models showed positive correlations with the social ostracism norm, $r_{\text{Claude 3.5}} = 0.36$, $p < .001$; $r_{\text{GPT-4}} = 0.04$, $p = .025$, while only Claude showed a significant positive correlation with verbal confrontation, $r_{\text{Claude 3.5}} = 0.10$, $p < .001$; $r_{\text{GPT-4}} = -0.01$, $p = .810$. In beta-regression models, only Claude showed a significant effect for social ostracism, $b = 0.26$, $p = .005$, $\text{OR}(95\%) = 1.29$ (1.08, 1.54), while GPT-4 did not, $b = 0.00$, $p = .897$, $\text{OR}(95\%) = 1.00$ (0.96, 1.05). Neither model showed significant effects for verbal confrontation, Claude: $b = 0.12$, $p = .141$, $\text{OR}(95\%) = 1.13$ (0.96, 1.34); GPT-4: $b = 0.00$, $p = .993$, $\text{OR}(95\%) = 1.00$ (0.97, 1.03). Claude's OR suggests a 29% increase in reported probability of self-censorship when the social ostracism norm increases by 1SD.

Figure 4

The correlation of the mean probabilities of self-censorship by nation, with the social exclusion meta-norm of the nation, split by LLM.

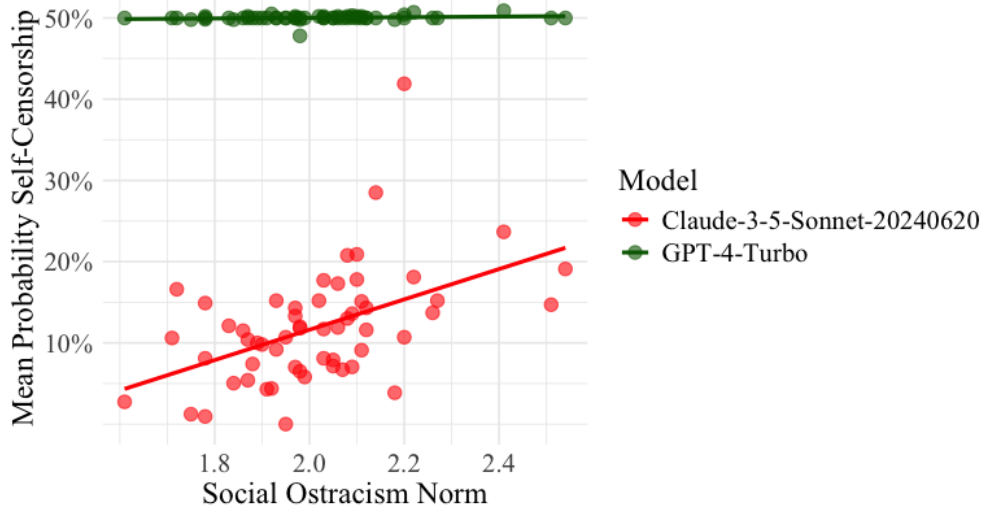


Table 3. Correlations between behavioral response probabilities and cultural meta-norms by model, and regression coefficients and odds ratios for the regression models with errors clustered by nation.

Dependent Variable	Claude 3.5 Sonnet		GPT-4	
	Social Ostracism	Verbal Confrontation	Social Ostracism	Verbal Confrontation
Ignoring	.16***	.04	.09***	-.05*
Downvoting	-.35***	-.03	-.10***	-.01
Reporting	-.03	.02	-.04	.00
Confrontation	-.28***	.08***	-.04*	.02
Self-censorship	.36***	.10***	.04*	-.01
Mean $ r $	.24	.054	.062	.018

	Social Ostracism	Verbal Confrontation	Social Ostracism	Verbal Confrontation
Ignoring	b = 0.06*, OR = 1.06 [1.01, 1.12]	b = 0.02, OR = 1.01 [0.97, 1.07]	b = 0.03*, OR = 1.02 [1.01, 1.05]	b = -0.01, OR = 0.99 [0.97, 1.02]
Downvoting	b = -0.25***, OR = 0.77 [0.69, 0.87]	b = -0.05, OR = 0.95 [0.85, 1.06]	b = -0.03***, OR = 0.97 [0.96, 0.99]	b = -0.01, OR = 0.99 [0.97, 1.02]
Reporting	b = -0.04, OR = 0.96 [0.88, 1.05]	b = 0.01, OR = 1.01 [0.94, 1.09]	b = -0.01, OR = 0.99 [0.98, 1.00]	b = 0.00, OR = 1.00 [0.99, 1.01]
Confrontation	b = -0.16*, OR = 0.85 [0.72, 1.00]	b = 0.03, OR = 1.04 [0.93, 1.15]	b = -0.01, OR = 0.99 [0.96, 1.00]	b = 0.01, OR = 1.01 [0.98, 1.03]
Self-censorship	b = 0.26**, OR = 1.29 [1.08, 1.54]	b = 0.12, OR = 1.13 [0.96, 1.34]	b = 0.00, OR = 1.00 [0.96, 1.05]	b = 0.00, OR = 1.00 [0.97, 1.03]

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.

2.4. Discussion

In Study 1, we investigated how LLM-simulations of the behavior of Reddit-users from different nationalities correlated with the respective meta-norms in those nationalities, specifically focusing on social ostracism norms and verbal confrontation norms. Overall, social ostracism norms correlated more strongly with most norm-enforcement behaviors than verbal confrontation norms. Moreover, probability-self-reports simulated with Claude 3.5 Sonnet correlated more strongly with the nations meta-norms than self-reports simulated with GPT-4-Turbo. Social ostracism norms were positively correlated with ignoring posts and self-censorship, and negatively

correlated with downvoting and confrontation. When controlling for the clustered error structure, all impact of the verbal confrontation norm became insignificant, while those for social ostracism held, though more reliably for Claude 3.5 Sonnet than for GPT-4-Turbo. The only behavior that did not seem to be related to at least one of the two meta-norms was *reporting* the post.

While some of these results are superficially intuitive, like social ostracism norms' positive correlation with the likelihood of ignoring an improper post and negative correlation with confronting the author others could plausibly have gone either way. The (rather strong) negative correlation of social ostracism norms with downvoting is one such case, as downvotes effectively enable exclusion of content via diminished reach. As for model differences, our validation study provided evidence that Claude 3.5 Sonnet proved better at modeling real world behaviors than GPT-4-Turbo, particularly when modeling social ostracism norm related behaviors. To a large extent, GPT-4-Turbo's deficiencies in this regard seem to stem from a strong underestimation of between-nation variance. Generally, the weaker correlations of the verbal confrontation norms could realistically have a similar origin, as *active* verbally confrontative behavior could collide with the models' policy and guidelines more strongly than *passive* socially ostracizing behavior. As role-playing is a common strategy to circumvent or "hack" the models (Toyer et al., 2023), this would also explain why they did not want to "play the character" that confronts either. Thus, as evidenced by the validation study, intercultural differences in verbal confrontation might be suppressed more strongly by the algorithm behind the models to either prevent harmful interaction or to avoid stereotyping socially undesirable behaviors.

In summary, the first study suggests that intercultural meta-norms *do* shape online discourse: Simulated users from nations with stronger social ostracism norms tended to remain passive insofar that they reported a higher likelihood of ignoring the comment and a lower likelihood for engaging in active punishments like confrontation or downvoting. Furthermore, the simulated respondents from nations with a stronger ostracism norm seemed more sensitive to indirect feedback, with a higher likelihood being reported to delete a post that "is not getting positive feedback". The correlations for the verbal confrontation norm were generally much weaker. Although these findings are interesting and of value by themselves already, nonetheless they merely represent a simulation of actual online discourses. Within Study 2, we thus examine

actual behavioral patterns of online discourse on Reddit, and how it is influenced by intercultural meta-norms.

3. Study 2

Study 2 extends Study 1 by collecting data from actual Reddit users, going beyond the simulation of user behavior. The aim of the study was to verify and solidify the findings from Study 1. Beyond grounding findings in real world data, this also allows us to go beyond other LLM-specific limitations on norm-enforcement behavior. It could thus be said that we are 1) interested in verifying and solidifying the effect of social ostracism meta-norms that was suggested by the LLMs, while 2) also reexamining a potential impact of verbal confrontation norms, as we have reason to believe that those norms were not sufficiently represented - or suppressed - in the LLMs. In total, Study 2 will not just give us a fuller picture on the impact of meta-norms on online norm-enforcement, but also on LLM-simulations as a research tool.

3.1. Method

3.1.1. Social Media Data

The dataset used in this study was downloaded from AcademicTorrents.com (AcademicTorrents, n.d.), a platform where researchers legally share large datasets using torrenting technology. The torrent contained all Reddit posts and comments from October 2021 to September 2023. In a quasi-experimental approach, we collected data from the nation-specific Subreddits, corresponding to the nationalities investigated by Eriksson et al. (2021). We gathered data for 45 out of the 57 total nations, only excluding nation-specific Subreddits that had fewer than 50,000 members – a threshold set to ensure a minimum of continuous activity within the forum. Several exclusion criteria were implemented to ensure data quality and focus on controversial posts. In the selection of posts, we excluded meta threads, which typically contain general discussions about the Subreddit itself or administrative announcements, from the analysis. To maintain a focus on substantive interactions, posts with fewer than one comment were omitted. Conversely, to limit extreme cases and concentrate on typical interactions, posts exceeding 500 comments were also excluded. To further refine the sample and eliminate outliers, posts with upvote ratios lower than 15% or higher than 75% were excluded. Additionally, posts with a score

of exactly 0.0 (receiving equally many upvotes and downvotes) were omitted from the analysis. To prevent the potential skewing of results by extremely popular posts, those with more than 500 upvotes or downvotes were excluded from the study. These criteria were designed to focus the analysis on a representative sample of controversial posts while maintaining data integrity. These exclusion criteria were pre-registered (see https://osf.io/38qf6/?view_only=8b855b69b6df4d49b481971b3b1e0a7d). For each of the first eleven Subreddits we sampled, exactly 5,000 posts were randomly sampled, resulting in a total of 45,000 posts (we had originally planned to contrast extremes of the two meta-norm distributions). After pivoting to a more comprehensive analysis, we sampled another 500 posts for the remaining nationalities, leading to a total of 62,504 posts. We then collected up to 10 comments per post, for a total of and 450,436 comments across all Subreddits. For information on the distribution of posts and comments across nationalities, as well as more details on exclusion criteria, see the OSF (Study 2 – Supplement 1b).

Lastly, we also collected the rules for all the Subreddits that were analyzed, to investigate whether they would also vary systematically according to meta-norms. Specifically, we only captured the headers for each of the rules of the Subreddits – e.g. “1) Be civil and follow the Reddiquette 2) Submissions must be on topic and relevant to Algeria 3) Use descriptive, non-editorialized and properly formatted titles 4) No low-quality content (...)”. This information was crucial to investigate potential causal paths of how intercultural meta-norms influence online norm-enforcement. More specifically, systematic differences in the content of the Subreddit-rules would support a potential explicit pathway, mediated by systematic differences in Subreddit-rules (see General Discussion).

3.1.2. Measures

In the following, we define a *post* as a novel contribution that is not explicitly posted in response to another contribution. In contrast, a *comment* is defined as a direct response to a previous contribution.

3.1.2.1. Share of negative comments

We conducted a sentiment analysis with GPT-3.5-Turbo (see Rathje et al., 2023), to categorize whether the comments were negative, in the sense of being a “remark or statement that expresses disapproval, criticism, or pessimism” (for the full prompt, see Study 2 – Supplement 1c). The “share of negative comments” was consequently the share of comments under a post that was negative. As comments with a score of exactly 0.0 were omitted from the analysis (see 3.1.1. Social Media Data) scores thus fell in a range from 0.1 (one out of the 10 sampled comments was negative) to 1 (all comments were negative).

3.1.2.2. Any negative comments

As many posts had no negative comments at all, we also included a dichotomous variable on whether there was any negative comment under a post (*yes* vs. *no*). For the model specification, we consequently split the distribution between a discrete and a continuous non-zero model.

3.1.2.3. Comment downvoted

Since there is no information on the explicit number of upvotes and downvotes for each comment, the only way to be sure whether a comment was downvoted was if the post had a negative score - i.e., more downvotes than upvotes. Thus, *comment downvoted* is “yes” when the score of the variable is below 0 (i.e., negative), and “no” in any other case.

3.1.2.4. Post was deleted

Whether a post was deleted by the author (*yes* vs. *no*).

3.1.2.5. Comment was deleted

Whether a comment was deleted by its author (*yes* vs. *no*).

3.1.2.6. Comment was removed

Whether a comment was removed by a moderator (*yes* vs. *no*) - A moderator on Reddit is a user who oversees a specific Subreddit, tasked with rule enforcement and community management (Reddit, 2024a).

3.1.2.7. Embedding of Subreddit-Rules

We translated the rules for each Subreddit to an embedding vector with 50 dimensions, translated by OpenAI's "text-embedding-3-large", which encodes a 3,072-dimensional representation of a text's semantic meaning. This allowed us to capture the meaning and content of the Rules at an abstract, statistically processible level (see Wang et al., 2024a).

3.2. Results

We will classify the reported results into three broad sections: a) *negative feedback* - which considers the prevalence of visible negative signals towards a post or comment, b) *censorship* - which considers both self-censorship, as well as censorship by moderators, and c) *Subreddit rules* - where we relate the content of Subreddit rules with meta-norms.

3.2.1. Negative Feedback

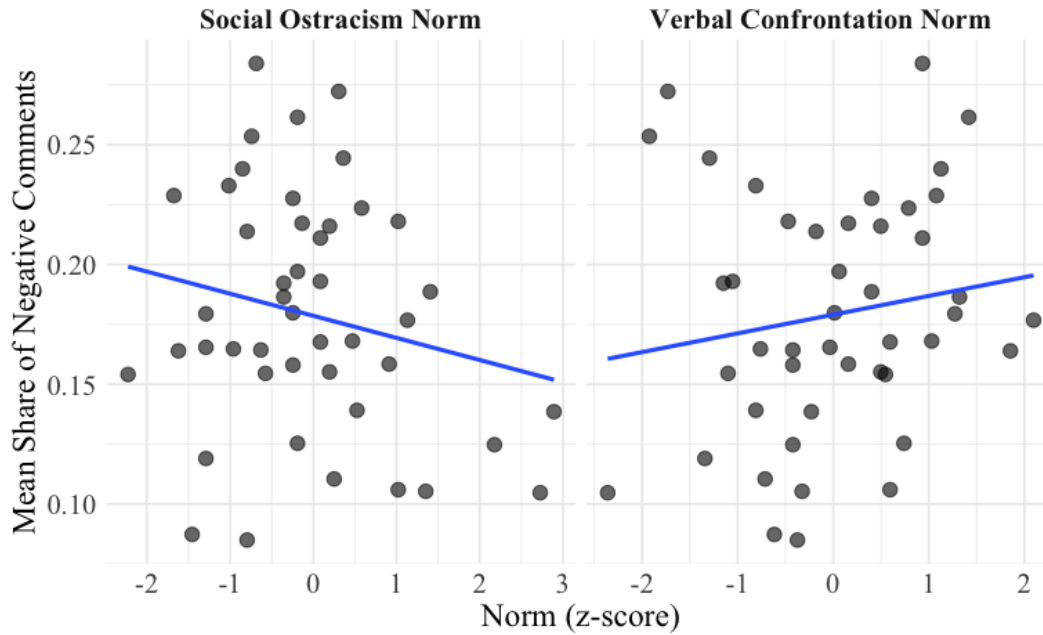
3.2.1.1. Likelihood and share of negative comments

We found a negative relationship between the strength of the social ostracism norm in the nation of a Subreddit and the likelihood of any negative comment on a post in that Subreddit, $b = -0.15$, $SE = 0.04$, $p < .001$, $OR (95\%-CI) = 0.86 (0.80, 0.92)$. This indicates that for one SD increase in the social ostracism norm, the odds of having any negative comment under a post decrease by 14%. There was no significant impact of the strength of the verbal confrontation norm on the likelihood of any negative comment on a post in that Subreddit, $b = -0.00$, $SE = 0.08$, $p = .989$, $OR (95\%-CI) = 1.00 (0.85, 1.17)$.

Among the nonzero shares, we also found a negative relationship between the strength of the social ostracism norm in the nation of a Subreddit and the share of negative comments on a post in that Subreddit, $b = -0.06$, $SE = 0.03$, $p = .039$. To illustrate this effect, the average share of negative comments within our dataset is 30.9% for posts with at least one negative comment. For a Subreddit with this average share, a one-unit increase in the strength of the social ostracism norm is associated with the share of negative comments decreasing to approximately 29.74%. This represents a decrease of about 1.16 percentage points or a relative decrease of 3.75%. Again, there was no significant impact of the strength of the verbal confrontation norm on the share of negative comments on a post in that Subreddit, $b = 0.00$, $SE = 0.03$, $p = .747$.

Figure 5

The mean share of negative comments in a nation-specific Subreddit, by the social ostracism norm and verbal confrontation norm in the nation of the Subreddit.

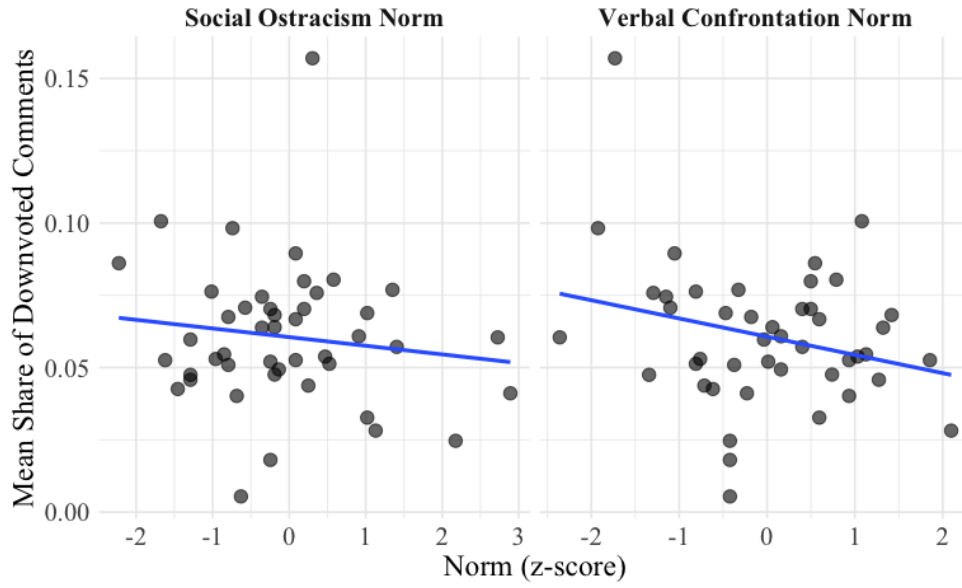


3.2.1.2. Share of downvoted comments

We found a negative relationship between the strength of the social ostracism norm in the nation of a Subreddit and the likelihood of a comment being downvoted in that Subreddit, $b = -0.17$, $SE = 0.05$, $p = .001$, $OR (95\%-CI) = 0.85 (0.76, 0.94)$. This indicates that a one SD increase in the strength of the social ostracism norm the odds of a comment being downvoted decrease by 15%. Similarly, there was a negative impact of the strength of the verbal confrontation norm in a nation and the likelihood of a comment being downvoted in that nation, $b = -0.21$, $SE = 0.08$, $p = .011$, $OR (95\%-CI) = 0.80 (0.69, 0.95)$, indicating that with a one SD increase in verbal confrontation norm, the likelihood of a post being downvoted decrease by 20%.

Figure 6

The mean share of downvoted comments in a nation-specific Subreddit, by the social ostracism norm and verbal confrontation norm in the nation of the Subreddit.



3.2.2. Censorship

3.2.2.1. Self-Censorship

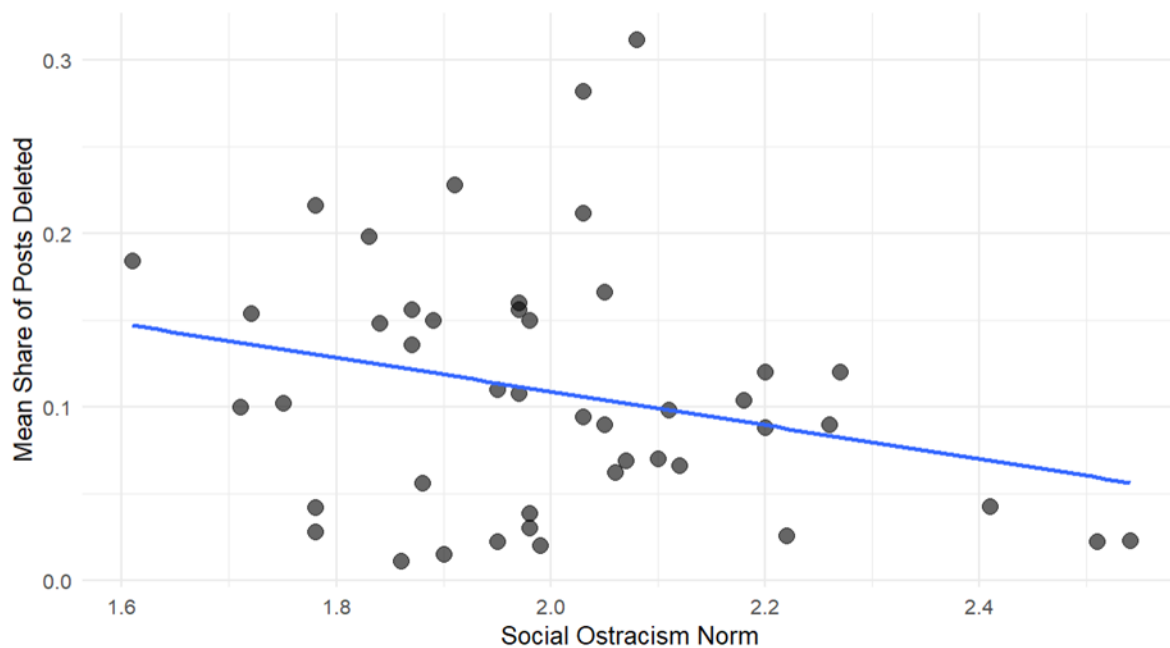
For *comments*, there was no significant relationship between the strength of the social ostracism norm in the nation of a Subreddit and the likelihood of the comment having been deleted by its author, $b = -0.09$, $SE = 0.05$, $p = .062$, OR (95%-CI) = 0.91 (0.83, 1.00). For the strength of the verbal confrontation norm on the other hand, the likelihood of an author deleting their *comment* was significantly negative, $b = -0.14$, $SE = 0.14$, $p < .001$, OR (95%-CI) = 0.87 (0.80, 0.94). This indicates that in Subreddits from nationalities with stronger verbal confrontation norms, authors are less likely to delete their comments. Specifically, for each one-standard-deviation increase in the strength of the verbal confrontation norm, the odds of an author deleting their comment decrease by 13%.

For *posts*, there was a significant negative relationship between the strength of the social ostracism norm in the nation of a Subreddit and the likelihood of the comment having been deleted by its author, $b = -0.35$, $SE = 0.08$, $p < .001$, OR (95%-CI) = 0.70 (0.60, 0.83). This indicates that in Subreddits from nationalities with stronger social ostracism norms, authors are less likely to delete their posts. Specifically, for each one-standard-deviation increase in the strength of the

verbal confrontation norm, the odds of an author deleting their comment decrease by 30%. For the strength of the verbal confrontation norm on the other hand, the likelihood of an author deleting their post was not significant, $b = -0.16$, $SE = 0.12$, $p = .173$, $OR (95\%-CI) = 0.85 (0.68, 1.07)$.

Figure 7

The mean share of posts deleted by users in a nation-specific Subreddit by the social ostracism norm.



By moderators. For comments, there was no significant relationship between the strength of the social ostracism norm in the nation of a Subreddit and the likelihood of a *comment* having been removed by a moderator, $b = -0.29$, $SE = 0.15$, $p = .055$, $OR (95\%-CI) = 0.75 (0.56, 1.01)$. For the strength of the verbal confrontation norm on the other hand, the likelihood of an author deleting their *comment* was significantly negative, $b = -0.56$, $SE = 0.12$, $p < .001$, $OR (95\%-CI) = 0.57 (0.45, 0.73)$. This indicates that in Subreddits from nationalities with stronger verbal confrontation norms, authors are less likely to delete their comments. Specifically, for each one-standard-deviation increase in the strength of the verbal confrontation norm, the odds of an author deleting their comment decrease by 43%.

3.2.3. *Meta-norms and Subreddit rules*

To determine whether the rules would vary systematically with the meta-norms, we calculated a distance correlation (Székely et al., 2007) between each meta-norm and the embeddings extracted from the rules of the respective Subreddits. In short, the relative distance between all values were quantified within each variable, and these relative distances were then correlated between the two variables – i.e., the embeddings of the vectors and each meta-norm, by nation. As a significance test, we tested how likely it was to achieve a correlation of this magnitude if the distance-array for one of the variables was randomly shuffled, also known as a permutation test (Hemerik et al., 2019). For the relationship between the social ostracism norm and the Subreddit-rules’ embeddings in a nation, a distance correlation test revealed a moderate but non-significant association between the strength of the social ostracism norm (?) in a nation and the semantic content, $dCor = .47$, $p = .075$, based on 1000 permutations. For the relationship between the verbal confrontation norm and the Subreddit-rules’ embeddings in a nation, a distance correlation test revealed a significant moderate association between the strength of the verbal confrontation norm in a nation and the semantic content, $dCor = .47$, $p = .022$, based on 1000 permutations.

3.3. *Discussion*

In Study 2, we aimed to 1) verify and solidify findings from the simulated data in Study 1, as well as 2) go beyond potential policy-based limitations of the LLM-simulations, especially regarding verbal confrontation norms.

Concurrent with the findings from Study 1, stronger social ostracism norms were associated with a decreased likelihood of confrontation, here coded as both the general likelihood as well as the share of negative comments (i.e., a comment that “expresses disapproval, criticism, or pessimism”). The negative relation between the strength of social ostracism norms and downvoting behavior was also replicated for the downvoting of comments. The finding from Study 1 that indicated a positive relationship between the strength of the social ostracism norm and self-censorship of posts did not replicate within Study 2. Concurrent with Claude 3.5 Sonnet’s findings in Study 1, Study 2 indicated a positive relationship between the strength of verbal confrontation norms and the likelihood of confrontation. Additionally, for verbal confrontation norms, the real-life data in Study 2 indicated several relations that were absent from the simulation data within

Study 1. Particularly, the strength of the verbal confrontation norm were negatively related to downvoting of comments. Moreover, comments in nationalities with stronger verbal confrontation norms were censored less, whether by the users themselves or by moderators. These effects of the separate meta-norms were independent of each other, as the models controlled for the impact of the respective other meta-norms. Interestingly, the independence is further emphasized by the incongruent impacts that the two norms have on different variables. For downvoting behavior, the effects of the two meta-norms align (i.e., both stronger social ostracism and stronger verbal confrontation norms being related to less downvoting behavior), while they are opposite for confrontation (i.e., stronger social ostracism norms being related to less confrontation behavior and stronger verbal confrontation norms being related to more confrontation behavior). Aside from the overall differences on the verbal confrontation norms, the main inconsistencies between the studies concerned censorship behavior. Specifically, relations of the social ostracism norms were opposite of what the LLMs had predicted, with stronger social ostracism norms being associated with more self-censorship in the LLM data and less self-censorship in the real life data. There are plenty of plausible explanations for this, one likely candidate being that self-censorship largely happens *before* a post is made (e.g., Das & Kramer, 2013) and that people in cultures with weaker social ostracism norms would risk less with a daring post - thus reducing the average share of daring posts. Yet, further studies would be needed to test this. Lastly, there was a relationship between the meta-norms in a nation and the embeddings (i.e., the semantic content) of the Subreddit rules. In other words, nationalities with more similar meta-norms also had more similar Subreddit rules, semantically, outlining one plausible causal path of how intercultural differences manifest online.

4. General Discussion

Online forums are becoming an increasingly important medium of social interaction around the globe (e.g., Auxier & Anderson, 2021). We investigated whether and how norm-enforcement in online forums is influenced by intercultural meta-norms, that is, norms on appropriate punishment/ norm-enforcement. We focused on two meta-norms specifically, namely social ostracism and verbal confrontation. To investigate our research question, we relied on experimental data from LLM-based simulations (Hewitt et al., 2024; Manning et al., 2024) and observational data from Reddit (Proferes et al., 2021). Both studies show that the norm-

enforcement strategies that are chosen in online forums are influenced by a nation's meta-norms. Particularly, the results indicated that in nationalities with higher social ostracism meta-norms, active forms of punishment are less likely, even if those forms of punishment functionally exclude posts (i.e., downvotes on Reddit). For nationalities with stronger verbal confrontation, results were not consistent across studies. Yet, as a validation pre-study had indicated, intercultural differences in verbal confrontation norms might not well represented in the LLMs, which might explain the non-significant findings for verbal confrontation norms in Study 1. In the observational Study 2 that relied on real-life data from actual reddit users, stronger verbal confrontation norms were related to more confrontation and less downvoting, as well as less overall censorship both by the authors of posts and comments themselves as well as moderators of the respective Subreddit.

4.1. Intercultural differences in norm-enforcement in online forums

The presented results suggest a relation between intercultural meta-norms and respective norm-enforcement in online forums. These effects were found, despite internet users plausibly having a general bias towards WEIRD (meta-) norms across cultures, as internet users are more likely to be Educated, Industrialized, and at least relatively Richer (though not more western, OECD, 2024) which naturally limits the variance. Having established intercultural differences in online norm-enforcement, a next step would be to examine the mechanisms driving these variations. A key question is whether these differences stem primarily from explicit rules that vary between nation-specific Subreddits (henceforth the *explicit path*), or from participants' implicit cultural meta-norms (henceforth the *implicit path*). Let's first consider the arguments for the explicit path. In Study 2, we found that there were moderately strong correlations between the embeddings of a nation's Subreddit-Rules and a nation's verbal confrontation and social ostracism norms (though the correlation was only significant for the verbal confrontation norm). This suggests that the semantic content of a Subreddit's rules varied systematically according to the nation's meta-norms. This is in line with findings from Kenntemich and Rudert (2024), who found a causal impact of "codes of conduct" in (artificial) online forums on the norm-enforcement behavior. While in combination, it seems highly likely that the explicit path explains at least a part of the overall effect, we cannot be sure whether the rules have a causal impact on the Subreddit's participants, or merely reflect the implicit meta-norm socialization of a nation's citizens, which would thus represent a shared covariate between behavior and rules. While our first study was

experimental, 1) simulation data alone provides insufficient evidence to determine real-world causality (see 4.2.), and 2) it seems entirely plausible that the LLMs deduce different contexts (here Subreddit-rules) into the context of the response. As for the implicit path, the depth of the literature on how cultural norms are internalized (e.g., Henrich et al., 2010a, 2010b; Marlowe et al., 2008) would suggest that explicit rules are not necessary to follow the rules that have been learned in the offline-context and thus simply feel like the appropriate thing to do, independent of context (Rietveld, 2008). Yet, a study by Qiu and colleagues (2013) suggests that individuals are at least partly aware of differences in cultural norms. In their study, they show that bicultural individuals using both Facebook (U.S.) and Renren (China) adapted their engagement styles based on platform norms. Particularly, they indicated a tendency for more cooperative, in-group sharing behaviors on Renren – a collectivist approach that was also observed by Hara and colleagues (2010, see 1.2.), whereas on Facebook, behaviors aligned more with individualistic self-presentation. This matches the findings of research on bilingual individuals, who report shifts in personality, according to the language they communicate in (Chen & Bond, 2010; Ramírez-Esparza et al., 2006), a phenomenon called “cultural frame switching”, here, cued by language. However, in the context of our study, the cultural frame switching could not solely have come from the different languages, as a majority of the Subreddits (36/57 - 63%) used English as their main language. Nonetheless, these findings on cultural frame switching raise important questions that should be the focus of future research.

4.1.1. Limitations and future perspectives on intercultural meta-norms online

To investigate intercultural differences, we decided to focus on nation specific Subreddits. As there is no information on user-nationality available, this was – from our perspective – the most fruitful approach to investigate intercultural differences. But it came with the obvious downside, that we focused on Subreddits that were rather culturally homogenous. In other words, people interacting in the context of the Subreddits had an obvious signal for the appropriate cultural frame (e.g., Oddný et al., 2023), thus potentially making the findings on inter-cultural differences rather context-, or rather frame-specific. Currently, the biggest Subreddits with international participation (e.g., r/worldnews or r/AskReddit) have English as the default language, and despite international participation, are dominated by the English-speaking and more broadly WEIRD participants (World Population Review, 2024b). While Study 2 might (purposefully) have tapped into the more

multicultural forums, this could also have introduced an unplanned source of variance. Hara and colleagues (2010) had identified community size itself as a predictor of qualitatively different interaction, with smaller communities, amongst others, showing *more conflict*. Future studies should thus aim to keep the characteristics of forums as constant as possible, investigating whether intercultural differences in norm-enforcement hold in culturally diverse settings. This question will likely become increasingly more relevant, as the dependence on English as the international language could decrease, as generative AI is steadily decreasing the inter-cultural language barriers (Vaswani et al., 2017), with tools like the Browser “Google Chrome” allowing for instant translation of a webpages content (Google, n.d.). While this is currently still limited to two languages at a time, truly international communities, unbound by language, will likely become increasingly seamless, while shedding the cultural frame previously provided by language (Chen & Bond, 2010; Ramírez-Esparza et al., 2006). As online communities become more language-fluid, they provide a unique window into how people adapt their behavior to different cultural contexts, independent of the normative guidelines that the English language carries, as a cultural frame – potentially allowing for insights on human social functioning that could extend beyond online interactions. The technical capabilities for studying this phenomenon already exist, independent of a potential widespread intercultural exchange that could follow from social media platforms fully integrating translation features. We by no means say that all online communities will converge to a global online community, just that such communities, unbound by language constraints and language’s cultural context, is now becoming possible. Whether such communities will become more prevalent is an entirely separate question, as adapting multicultural identities could come with its own set of challenges (e.g. Boros et al., 2019; Sparrow, 2000).

4.2. Methodological Implications

4.2.1. Using Large-Language Models to simulate intercultural meta-norms

Despite gaining an increasing amount of traction recently (e.g. Bail, 2023; Dillion et al., 2023) using generative AI in the social sciences is still a novel approach, and is worth reflecting on, for the current use-case. 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). Furthermore, simulating conflict behavior could be a particularly interesting use-case, since this behavior could collide with the models’ guidelines for appropriate behavior (Anthropic, 2024; OpenAI, 2024). Lastly, here, we also compared the outcomes for two separate frontier models, in GPT-4-Turbo and Claude 3.5 Sonnet (June 2024 release), an approach that has also been less common in social scientific research.

4.2.1.1. Intercultural differences

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. Our simulation study (Study 1) indicates a mixed picture: While the models were highly accurate at predicting reactions to particular situations in general (see 2.1. *Pre-Study: Validation of the LLM-simulation approach*), particularly GPT-4-Turbo did a rather poor job of simulating intercultural differences in reported behavior. While we decided to prompt all models in English, which arguably could have decreased the power of language in triggering intercultural representations, Claude 3.5 Sonnet frequently (32% of all responses) responded in non-English languages when assuming the role anyway, which GPT-4-Turbo did not do. This assumption of languages by Claude 3.5 Sonnet was particularly frequent in Spanish- and Portuguese-speaking nationalities (e.g., Brazil, Spain, Mexico, Mozambique) and Italy, for whom most responses were non-English (83.8%). Even English-proficiency seems to somewhat have been taken into account here, as Portugal, despite its obvious connection to Portuguese, nonetheless responded in English most often of the Spanish-Portuguese language cluster, which could have reflected the typically high English-proficiency of Portuguese (World Population Review, 2024a). But after these outliers, Vietnam (24%) had the next-most non-English responses and 53% of all nationalities, that did not have English as their native language (e.g., including China, Saudi Arabia, and Greece), had no non-English responses at all.

Overall, LLMs seem to have the potential to map intercultural differences, potentially, with improved predictive accuracy when they adopt the linguistic patterns characteristic of a specific culture.

4.2.1.2. Model differences and intransparencies

Claude 3.5. Sonnet being better at encoding intercultural variance in norm-enforcement behaviors is in itself an interesting and noteworthy finding, as it introduces another dimension into LLM-based social scientific research – (frontier) model selection. While Claude 3.5 Sonnet outperforms most other models on a wide range of benchmarks (Anthropic, 2024) and might to some extent just be a better model, the model-alignment strategies of the two companies might also have played a role. The lack of variance in verbal confrontation norms seems, at least partially, to have been related to OpenAI’s harmful content policies. The style of interaction is learned in the post-training of the models (OpenAI, 2023b), which essentially means that these styles of interaction are not strictly learned from data and modeling language but are also guided by explicit policies that are set by the respective companies behind the models (Bai et al., 2022; Christiano et al., 2017). When asking the models to simulate a situation of (minor) conflict (Study 1), the effect of the post-training might have influenced what kinds of responses were considered probable. The importance of avoiding stereotypes in OpenAI’s policy is hinted at when asking the newly released “o1-preview” to assess the same response probabilities as were modeled in Study 1. As o1 models and provides a transparent (though summarized) “thought process”, the explicit step of “Avoiding harmful stereotypes” can be traced throughout the reasoning process (see Study 1 – Supplement 1a). Thus, the differences between the models could be due to a different weighing of intercultural neutrality, or even a generally stronger focus to date of Anthropic on avoiding *false refusals* (i.e., incorrectly classifying requests as harmful; Anthropic, 2024).

We agree that a critical perspective on modeling human behavior with LLMs is appropriate, and that it should not be a sole source of evidence (e.g., Agnew et al., 2024). Yet, we also believe that the comprehensive, if flawed world model captured in LLMs is a valuable resource for exploratory social science research, especially as our understanding of these models increases (Templeton, 2024; Binz & Schulz, 2023), and when combined with more traditional approaches. Recent efforts have even aimed at fine-tuning LLMs for the particular purpose of modelling humans, which could further increase the merits of LLMs as a cost-effective tool for initial exploration of human behavior (Binz et al., 2024).

4.2.2. Using vector-embeddings as a measure for semantic similarity of texts

We also used vector embeddings of the Subreddits' rules (Study 2) to encode high-dimensional semantic representations of the rules (Palangi et al., 2015). This allowed us to quantify similarities between these semantic representations, and even to correlate the vector embeddings with the meta-norms, as two data points per nation. Generally, we consider vector embeddings to be an innovative and powerful tool for text data analysis, especially when 1) an abstract representation of semantic meaning is enough (as was the case here), or 2) any concrete categorization is approached inductively (i.e., a bottom-up discovery process), for example by first clustering text input, and then labeling the clusters post-hoc. As a simplified explanation, embeddings are like the first step of what LLMs like GPT-4-Turbo, when generating text-responses, it is their medium of “understanding” language. Basically, working with embeddings is like letting an LLM only encode the text, without further processing the encoded input, to generate an appropriate next word. With a maximum input of 8,191 tokens (~5,500 words), OpenAI's embedding models (OpenAI, n.d.) are a highly flexible (and cost-effective) tool for text data analysis.

5. Conclusion

In two studies, one with LLM-based simulations and one observational study of Reddit data, we found that norm-enforcement in online forums differs interculturally. Patterns of intercultural differences in perceptions of how norms *should* be enforced (i.e., Meta-norms), correlated with particular behavioral patterns of online norm-enforcement. Simulated and real Reddit users from nationalities with stronger social ostracism norms, were less likely to show active online norm-enforcement, like posting a critical comment or downvoting. Real Reddit users from nationalities with higher verbal confrontation norms, were more likely to post negative comments, though also less likely to downvote.

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.

Appendix A. Supplementary data and materials

Supplementary data and materials to this article can be found online at https://osf.io/uzdty/?view_only=6abce2111a3942a8b65438779316ab8b

References

- AcademicTorrents. (n.d.). *AcademicTorrents.com*. Retrieved from <https://academictorrents.com>.
August 11, 2024.
- Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2024)*. arXiv. <https://doi.org/10.48550/arXiv.2401.08572>
- Anthropic. (2024). *Claude 3 model card*. Anthropic. Retrieved from <https://docs.anthropic.com/en/docs/resources/model-card>. Accessed November 7, 2024.
- Auxier, B., & Anderson, M. (2021, April 7). 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. *The American Political Science Review*, 80(4), 1095-1111. <https://doi.org/10.2307/1960858>

- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*. <https://arxiv.org/abs/2204.05862>
- 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., & 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>
- 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>
- 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>

- 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, 107062. [10.1016/j.chb.2021.107062](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>
- Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, 15(5), 218–226. <https://doi.org/10.1016/j.tics.2011.03.003>
- Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017. <https://doi.org/10.48550/arXiv.1706.03741>
- 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.
- 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., Gächter, S. Altruistic punishment in humans. *Nature* 415, 137–140 (2002). <https://doi.org/10.1038/415137a>

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

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>

Gintis, H. (2003). The Hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms. *Journal of Theoretical Biology*, 220(4), 407–418. <https://doi.org/10.1006/jtbi.2003.3104>

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., & 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>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *The Behavioral and brain sciences*, 33(2-3), 61–135. <https://doi.org/10.1017/S0140525X0999152X>
- 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>
- 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.
- Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1014>
- 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 Acknowledgement-Seeking Behavior. *Personality & Social Psychology Bulletin*.
<https://doi.org/10.1177/01461672241297824>
- Kenntemich, C. & Rudert, S.C. (2024). 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>
- Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28(2), 434–443.
<https://doi.org/10.1016/j.chb.2011.10.014>
- Lapidot-Lefler, N., & Barak, A. (2015). The benign online disinhibition effect: Could situational factors induce self-disclosure and prosocial behaviors? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 9(2). <https://doi.org/10.5817/CP2015-2-3>
- 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, 104040. <https://doi.org/10.1016/j.cognition.2019.104040>
- Meta. (2022). *Keep building better: The Meta code of conduct*. Retrieved from https://s21.q4cdn.com/399680738/files/doc_downloads/governance_documents/2022/06/FB_CoC_EXTERNAL_en_EN_Update_Final-6_2-FINAL-ua.pdf. Accessed November 6, 2024.
- 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), 20230034. <https://doi.org/10.1098/rstb.2023.0034>
- OECD. (2024). *Internet use varies across educational attainment, age and household income: Internet users by gender, age, education, and household income, as a percentage in each group, OECD average 2022*. In *Society at a Glance 2024: OECD Social Indicators*. OECD Publishing. <https://doi.org/10.1787/918d8db3-en>
- 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. (n.d.). *Embedding models: Use cases*. OpenAI. Retrieved from <https://platform.openai.com/docs/guides/embeddings/use-cases#embedding-models>. Accessed November 7, 2024.
- OpenAI. (2023a). *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 November 7, 2024.

OpenAI. (2023b). *GPT-4 technical report*. Retrieved from <https://cdn.openai.com/papers/gpt-4.pdf>.

Accessed November 7, 2024.

OpenAI. (2024). *GPT-4 system card*. OpenAI. Retrieved from [https://cdn.openai.com/papers/gpt-4-](https://cdn.openai.com/papers/gpt-4-system-card.pdf)

[system-card.pdf](https://cdn.openai.com/papers/gpt-4-system-card.pdf). Accessed November 7, 2024.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., & Ward, R. (2016). Deep

Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4), 694–707. <https://doi.org/10.1109/TASLP.2016.2520371>

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7(2).

<https://doi.org/10.1177/20563051211019004>

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>

Rathje, S., Mirea, D.-M., Sucholutsky, I., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34), Article e2308950121. <https://doi.org/10.1073/pnas.2308950121>

- Reddit. (n.d.-a). *What's a moderator?* Reddit Help. Retrieved from <https://support.reddithelp.com/hc/en-us/articles/204533859-What-s-a-moderator>. Accessed November 5, 2024.
- 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 October 16, 2024.
- Reddit (2024b). *Reddit and OpenAI build partnership*. Reddit Announcements. Retrieved from <https://www.redditinc.com/blog/reddit-and-oai-partner>. Accessed November 5, 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., & Kenntemich, C. (2024). *The impact of restrictive codes of conduct on norm-enforcement in online forums*. Manuscript submitted for publication.
- Safari, Rahebeh Mojtahedi; Rahmani, Amir Masoud; Alizadeh, Sasan H. . (2019). *User behavior mining on social media: a systematic literature review*. *Multimedia Tools and Applications*, 33747–33804. doi:10.1007/s11042-019-08046-6
- Schramowski, P., Turan, C., Andersen, N. *et al.* Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat Mach Intell* 4, 258–268 (2022). <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>
- SimpleBench. (n.d.). *SimpleBench: Where everyday human reasoning still surpasses frontier models*. Retrieved from <https://simple-bench.com/>. Accessed November 5, 2024.
- Sparrow, Lise. (2000). Beyond multicultural man: Complexities of identity. *International Journal of Intercultural Relations*. 24. 173-201. 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/10.1214/0090536070000000505>
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Summers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., & Henighan, T. (2024). *Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet*. Anthropic.
- 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>
- Twitter. (2018). *Code of business conduct & ethics*. <https://cdn.cms-twdigitalassets.com/content/dam/legal-twitter/asset-download-files/en/code-of-conduct-092018-ext.pdf>

- 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. (2024). 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. (2024). *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>
- 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, 361–373. <https://doi.org/10.3758/s13428-014-0475-x>
- World Population Review. (2024a). *English proficiency by country*. Retrieved from <https://worldpopulationreview.com/country-rankings/english-proficiency-by-country>. Accessed November 7, 2024.
- World Population Review. (2024b). *Reddit users by country*. Retrieved from <https://worldpopulationreview.com/country-rankings/reddit-users-by-country>. Accessed November 7, 2024.

The impact of conflict-discouraging codes of conduct on norm-enforcement in online forums

Word Count: 13,986

Abstract

How does conflict-discouraging (vs. conflict-encouraging) code of conduct impact norm-enforcement behavior in online forums? In 5 experimental studies (Total N = 976), we investigate the impact of conflict-discouraging vs. conflict-encouraging on norm-enforcement behavior, specifically investigating rates of confrontation (i.e., active signaling of discontent) and ostracism (i.e., not responding) of norm-violating targets. Using a novel paradigm that simulates an online forum, we find that norm-violating targets are more likely to be ostracized when codes of conduct discourage conflict and more likely to be confronted when codes of conduct encourage conflict. There are no robust differences in the rates of inclusion, indicating that the code of conduct can shift the balance between the two tested strategies of norm-enforcement. Our findings were only robust when targets violated interaction norms (i.e., rude language), but not when targets violated social exchange norms (i.e., stole an idea). Consequences for guiding norm-enforcement behavior online are discussed.

1. Introduction

Social norms are the informal rules and expectations that guide group behavior (Bicchieri, 2006). They are a potent social technology that can coordinate interaction on large scales (Cialdini & Trost, 1998). Large-scale studies have shown that the explicit communication of expected and appropriate behavior - i.e., codes of conduct - can decrease conflict and unruly behavior both offline (Paluck et al., 2016) and online (Matias, 2019). But social behaviors can be complex, and suppressing an undesired behavior does not necessarily lead to a desired one (e.g., Hunter et al., 2019). While large-scale studies show a decrease in reported conflict and harassing comments respectively, could the antisocial pressure just have been released in a way that does not risk violating conflict-discouraging codes of conduct? One alternative behavior, for instance, could be to ostracize potential targets, withholding social interaction by ignoring or excluding targets (Robinson, 2013). Ostracism has been shown to be a preferred tool, when more direct confrontation could come at a greater social or personal cost for the source of ostracism (Williams & Nida, 2011), which might be the case for online forums characterized by explicit conflict-discouraging codes of conduct. Such a shift in punishment strategy could be especially worrisome when considering that social ostracism can be perceived as more aversive than even hostile acknowledgement (Rudert et al., 2017).

Here, we investigate both ostracism and confrontation as potential reactions to norm-violations in the context of online forums, and how their prevalence is influenced by the liberties that codes of conduct grant towards practices of conflict. Specifically, we propose that *conflict-discouraging* codes of conduct will increase the likelihood of ostracism in online forums, while decreasing the likelihood of confrontation, relative to *conflict-encouraging* codes of conduct.

1.1. Strategies of norm-enforcement

Norms are a crucial component for maintaining a cooperative social environment, from a societal level to small group settings, informing which behavior is considered normal and appropriate (Cialdini et al., 1991). Norms inform what behavior is considered *normal* in a certain context - known as *descriptive norms* - and they can inform what behavior is *expected* in a certain context - known as

injunctive norms (Cialdini & Trost, 1998). Consequently, deviation from what is normal might be surprising, but deviation from what is expected is inappropriate, and punished (Cialdini et al., 1991), to decrease unpredictable or exploitative behavior (Fehr & Fischbacher, 2003; Hales et al., 2017; Henrich et al., 2006). The specific targets of this organizing mechanism broadly target two goals: *interaction norms*, including politeness and respectful behavior, aim to simplify interactions by providing structure and predictability and 2) *generic exchange norms*, including reciprocity, loyalty and cooperation, aim to regulate social exchange and cooperation (Kerr & Levine, 2008). What behavior is expected in certain social contexts, can either be inferred from what behavior is punished (Abbink et al., 2017; Villatoro et al., 2014), but it can also be stated explicitly (e.g., Cialdini et al., 2006). In online forums both of these mechanisms are prevalent: Inferences on what is considered inappropriate can be made based on observed punishments (e.g. Munger, 2017), and explicit rules often communicate what kind of behavior is expected within an online forum (Chandrasekharan et al., 2018) - what we will henceforth refer to as *codes of conduct*.

Codes of conduct are thus a tool aiming to actively shape the social interaction processes in online forums. But here, the necessary interaction of different normative processes becomes apparent: The aforementioned observed punishments must be adhering to the explicit norms set by the code of conduct, or punishers risk becoming the punished. In the following we will outline two prominent strategies of norm-enforcement - confrontation and ostracism - before then considering how the attractiveness and prevalence of each strategy might be influenced by characteristics of the code of conduct.

1.1.1. Confrontation

Confrontation captures a variety of behaviors that all have in common that they *actively engage* a target and confront the target with one's disapproval about the perceived norm violation. Confronting behavior can take many shapes and forms, from a tactful and kind inquiry to change one's behavior to aggressive physical retaliation (cf. Del Carmen Herrera et al, 2018; Dickter & Newton, 2013; Feagin, 1991; Swim & Hyers, 1999).

In general, confrontation of norm-violating others has been associated with many desirable outcomes. Perhaps most importantly, confrontation effectively reduces future norm-violations in targets. A plethora of studies supports that confrontation is a highly effective way to reduce the likelihood of unwanted behavior in the future (Czopp et al., 2006; Mallett & Wagner, 2011; Montieth et al., 2002). Furthermore, this effectiveness of confrontation to prevent future norm-violations is robust, as it holds for hostile or friendly confrontations, as well as different sources of confrontation (i.e., victim vs. third-party observer) (Czopp et al., 2006; Drury & Kaiser, 2014). Additionally, the confronting person can benefit personally from their behavior, as confrontation can signal one's trustworthiness to observers (Jordan et al., 2016; Jordan & Rand, 2020) and increase prestige as well as dominance (Redhead et al., 2021). In the online forums, algorithms seem to particularly reward confrontation that is emotional and moralizing (Brady et al., 2023), revealing a related incentive for confrontation online and offline, given that attention is a proxy for prestige in online forums (Lu et al., 2012).

But there are also risks attached to these potential benefits. For one, there is the response of the confronted norm-violator, who could retaliate and punish a confronting person in return (Balafoutas et al., 2016; Nikiforakis, 2008; Rashidi, 2016), though this is likely far less threatening in anonymous online forums (Lapidot-Leifler & Barak, 2012). Nonetheless, confrontation can potentially entail social costs and can thus be a risky response to an observed norm-violation (Eriksson et al., 2017; Kaiser & Miller, 2001; Strimling & Eriksson, 2014): Particularly, observers of the confrontation can side with the confronted target and turn against the confronting person (Eriksson et al., 2017; Strimling & Eriksson, 2014). If these risks are considered too high by an observer of a norm-violation, an alternative and potentially lower-risk response is *ostracism*.

1.1.2. Ostracism

Ostracism is defined as excluding and ignoring others (Williams, 2007). In contrast to confrontations, which actively engage the target, ostracism is purposefully *omitting to engage* the target (Robinson, 2013). Like confrontation, ostracism is also a frequently used tool to punish and

correct norm-violating behavior (Rudert et al., 2020, 2023; Hales et al., 2017) and can promote cooperation in groups (Feinberg et al., 2014; Kerr et al., 2009; Maier-Rigaud et al., 2010). Ostracized individuals often seek to change their behavior, as ostracism has high social and emotional costs for its targets (e.g. Eisenberger et al., 2003; Rudert et al., 2021; Williams, 2007). These effects of ostracism transfer to online settings, as ostracism in online forums has caused similar emotional responses (Hayes et al., 2018; Wolf et al., 2015), as well as behavioral change in subsequent forum activity (Kenntemich et al., 2024b).

In contrast to confrontation, ostracism is less likely to elicit retaliation, as ostracism is more ambiguous and there is no direct contact with the target (Archer & Coyne, 2005). For third-parties, ostracism is harder to observe, as it represents, by definition, the *absence* of interaction behavior (Robinson, 2013). On the downside, in comparison to confrontation, ostracism could be less efficient at changing the target's behavior (cf. Molho et al., 2016). For one, unless the norm-violation was blatant and obvious even to the norm-violator, ostracism comes with more uncertainty for its target (Williams, 2009) and could even remain unnoticed. Therefore, it is less likely that a target of ostracism realizes that they offended others, compared to a target of confrontation. Especially in anonymous online forums, where not engaging is practically invisible (Kang, 2017; Paskuda & Lewkowicz, 2017), the attached risks for the individual are likely miniscule. While each individual's impact in ostracism is likely smaller, Kenntemich and colleagues (2024b) showed that the degree of behavioral consequences corresponds to the lack of positive feedback, making ostracism online a democratic and quantitative kind of norm enforcement.

In sum, both confrontation and ostracism are tools to sanction norm-violating behavior. While confrontation is typically more effective at promptly changing norm-violations, it also comes with more risks than ostracism. Thus, individuals are more likely to engage in confrontation when the associated risks are lower, but may revert to ostracism as a safer alternative when the risks of a confrontation are high. While several factors may influence the risks associated with confrontation - such as various factors inherent to the source (e.g. personally victimized or not – Molho et al., 2020), the target (e.g.

gender of target – Freedman et al. 2022), as well as their relationship (e.g. power hierarchy between source and target – Ashburn-Nardo et al., 2014) - here, we focus on codes of conduct. As stated earlier, codes of conduct are an explicit communication of what behavior is considered appropriate in an online forum (e.g. “Rules” in Subreddits, Reddit.com, 2024), offering a potential direct link into the cost calculation on choosing to confront vs. ostracize.

1.2. Codes of conduct and norm-enforcement

Not only norm-violators, but also observers are subjected to the social norms of their environment (Martin et al., 2019). Therefore, observers of a norm-violation must adhere to certain rules of behavior themselves *when responding to norm-violations*, to avoid the risk of becoming the target of subsequent punishment (e.g. Eriksson et al., 2017; Axelrod, 1986). In online forums, the aforementioned codes of conduct are an explicit normative tool to shape social interaction (Grace et al., 2022; Munger, 2017). Research by Kenntemich and colleagues (2024a), suggests a potential impact of codes of conduct on norm-enforcement practices on Reddit. Here, intercultural differences in meta-norms (i.e., norms on how to enforce norms) mapped onto differences in the codes of conduct, in country specific Subreddits, which in turn had systematically different patterns of norm-enforcement. In other words, the intercultural differences in the codes of conduct could be one potential mediator of intercultural differences in norm-enforcement practices on reddit. Yet, the observational design of these analyses does not allow for a causal interpretation.

As codes of conduct often aim to curb harassment within the forum (e.g. Munger, 2017), they often have to manage the slippery slope of where healthy conflict ends and where harassment begins (Grey, 1991). Here, different online forums often have differing philosophies, leading to differing levels in their acceptance of conflict.

1.3. Acceptance of conflict in online forums

As lack of social presence online, characterized by features like anonymity, invisibility, and lack of eye-contact amongst others, has been shown to markedly change the social behavior for the worse (Lapidot-Lefler & Barak, 2012), many online communities have fallen back on increasingly restrictive codes of conduct to avoid harassment. In a large-scale study of Reddit's moderation practices, Chandrasekharan et al.'s (2018) analyzed over 2.8 million removed comments across 100 subreddits, uncovering a complex ecosystem of norms that often show extreme caution towards conflict. For instance, in a cluster of 22 heavily moderated subreddits, moderators removed not only uncivil content but also "Personal reactions, opinions, and (failed) attempts to make jokes or be sarcastic" (Chandrasekharan et al., 2018, p. 17). Furthermore, in a large cluster of 43 diverse subreddits (including multiple forums with millions of users), moderators discouraged "criticizing other users' opinions" (Chandrasekharan et al., 2018, p. 17).

Such strict aversion towards conflict, could, however come with unintended consequences. In a comparative study of two Reddit forums with differing moderation policies, Gibson (2019) found that in the "safe space" subreddit, which had stricter moderation policies, there were significantly higher rates of comment removal by moderators and self-deletion by users compared to the "free speech" subreddit. Such increased levels of self-censorship demonstrates that users are more worried their activity might not be coherent with the forum norms. But measuring the deletion of previous posts might only be the tip of the iceberg of self-censorship. The risk of social sanction that is considered when deciding to delete a previous post might also be considered when deciding whether to comment at all. Thus, when witnessing inappropriate behavior in an online forum, the chosen norm-enforcement strategy might be dependent on how restrictive the forum's norms are towards conflict. Especially in a conflict-discouraging setting, like the "Safe Spaces" from Gibson (2019), ignoring other forum members as a subtle form of disapproval, might be the lower-risk - and thus preferred - alternative (Hales et al., 2017). Thus, we hypothesize that conflict-discouraging norms may have unintended consequences and increase the rate of ostracism, when a norm-violation is observed in an online-forum. In contrast, if the communicated codes of conduct encourage conflict - explicitly signaling users

that critical discourse is appreciated, the social risks of confrontation decrease drastically. Thus, while ostracism does not necessarily become a less attractive response, confrontation becomes more attractive, and thus more likely.

Taken together, we hypothesize that 1) ostracism is a plausible alternative to confrontation that is frequently used to subtly signal disapproval towards norm-violating behavior, and 2) the codes of conduct influence which of the two norm-enforcement strategies is chosen. Specifically, conflict-discouraging norms will lead to a higher likelihood of ostracism of norm-violating targets, compared to conflict-encouraging norms. Conflict-encouraging codes of conduct, on the other hand, will lead to a higher rate of confrontation of norm-violating targets, compared to both conflict-discouraging interaction.

1.4. Overview of the Studies

In the present studies, we pursue several aims: First, Studies 1 - 2 aim to introduce and test a paradigm that allows for both confrontation and ostracism as responses to an observed interaction norm-violation, in an online-forum-like setting. Study 3 introduces a manipulation of the code of conduct (henceforth CoC) as conflict-encouraging vs. conflict-discouraging (vs. none), to experimentally test their impact on the likelihood of ostracism, confrontation and inclusion in reaction to the *interaction norm-violation*. Studies 4 and 6 aim to replicate and generalize the findings from Study 3 to responses to violations of *generic exchange norms*.

1.4.1. Method Statement

1.4.1.1. Ethics Approval

All studies' procedures were approved by the University of [masked for peer review] Institutional Review Board and conform to the Declaration of Helsinki.

1.4.1.2. Open Science Practices

We report how we determined our sample sizes, all data exclusions, all data manipulations, and measures. Verbatim materials, analysis code, and data for all studies are available via the OSF (https://osf.io/nk5a8/?view_only=fc99ab6692e5431dbdf9069692cab6cd). We preregistered all hypotheses, designs, sample sizes, exclusion criteria, and analysis plans of the experimental studies on AsPredicted, (Study 1: https://aspredicted.org/ZWH_CRN; Study 2: https://aspredicted.org/VF9_4WR; Study 3: https://aspredicted.org/PNQ_952; Study 4: https://aspredicted.org/DMV_8T7; Study 5: https://aspredicted.org/F6P_6ZY).¹

1.4.1.3. Effect-size Measures

Throughout the manuscript, as effect size measures, we report Cohen's d for t-tests, bs for regression analyses, Cohen's w and Odds Ratios for Chi Squared tests, and Cramer's V for Fisher's exact test.

1.4.1.4. Text Coding

All text-categorization of messages that are reported, were coded with the Large-Language-Model Claude Sonnet 3.5 (see Rathje et al., 2023). The consistency between 267 coded messages that were coded by two human raters - including a coder-conference to clear up uncertain cases, and the Language-Model was good (Cohen's-Kappa = 0.75, $p < .001$). The prompt that instructed the model how to categorize the messages, can be found in the OSF (see *Data Analysis*).

1.4.1.5. Text-coded vs. Emoji-coded

We employed two strategies to code whether a message was a confrontation (vs. inclusion) of the target: The *emoji-coded* categorization was based on a choice of emoji that would supposedly capture what the participant wanted to convey with their message. Secondly, we coded the text content of the messages. As these two measures generally correlated highly, we will mostly only report the results of the emoji-based analyses, to avoid repetition. Where reporting both measures added

¹ We tested several exploratory predictors of norm-enforcement behavior. Particularly, Studies 1 and 2 tested whether predicted the chosen response, whereas Studies 3, 4, and 5 tested whether emotional reactions predicted the chosen response. In short, there were no robust predictive motives or emotions, and we therefore decided to cut them from the results to focus the narrative.

value to the contribution, we did so. Also, the analyses on the OSF always contain the analyses for both variables.

2. Study 1

Study 1 was to develop and test a new paradigm that would 1) realistically simulate interactions and norm-violations within an online setting, 2) allow for participants to react both with confrontation as well as ostracism to an observer's norm-violation, while 3) allowing for further experimental variations of the social context such as the prevailing injunctive social norms. We experimentally varied the behavior of the target (norm-conforming vs. norm-violating) and measured participants' behavioral response. Particularly, participants could choose between a) confronting the target by responding with displeasure with the target's behavior, b) including the target by responding with anything other than displeasure with the target's behavior, or c) ostracize the target by choosing to skip responding to the target and exclude them from all further interaction with the participant. We predicted that participants would engage more frequently in both ostracism and confrontation when interacting with a norm-violating target (vs. norm-conforming target). As this was the first experimental study to observe ostracism and confrontation simultaneously, the study also served the important purpose to determine a baseline proportion of the two behaviors in this paradigm, to build upon in later studies.

2.1. Method

2.1.1. Participants and Design

Participants were recruited from Prolific Academic (US Americans only) to take part in a study on "Interaction online", for a payment of GBP 1.25. Sample size was determined with GPower (Faul et al, 2007), according to which 141 participants provided sufficient power ($\approx .90$) to show a medium-size effect ($w = 0.3$, $df = 2$) of norm-violating (vs. norm-conforming) behavior on behavior choice (confrontation vs. ostracism vs. inclusion). To account for potential data exclusions, we collected data

from 162 participants. Two persons indicated that their data should not be used for analysis, thus, the final sample consisted of 160 US-Americans (108 female, 2 non-binary, $M_{\text{age}} = 34.1$, $SD = 12.1$) who were randomly assigned to two experimental conditions (target behavior: norm-violating vs. norm-consistent).

2.1.2. Materials and Procedure

In the newly developed paradigm, henceforth referred to as *Art Eval*, participants are confronted with the comments of alleged other participants about presented paintings and can choose how to respond to these comments. Participants were told that the study's aim was to investigate how experiencing art in an online community can shape the experience. After an instruction, participants chose a nickname for the interaction, and answered three questions about their interest in art. After that, they were presented with their first painting. The participants were asked to take a moment to look at the painting and then write a comment about their experience. After submitting their comment, participants were presented with comments of two alleged other participants, a control person and a (norm-conforming vs. norm-violating) target person. The comment of the control person was presented first and was norm-conforming, tactful, and positive, giving a detailed insight into the control person's experience of the painting. This served the dual purpose of increasing the plausibility of the experience and additionally enabled within-subject comparisons of responses. Next, participants were presented with the comment of the target: In the *norm-conforming* condition, the target criticized the painting in a civilized and polite manner (*"To me, the artist does not seem to be very skilled. The motive is kind of derivative and the craftsmanship seems to be rather fundamental, when using the right tools."*). In the *norm-violating* condition, the target criticized the painting in a rude and offensive manner (*"The painter seems to be a retard. I mean I could have drawn this shit in my sleep and that's his profession. I mean come on, what skill besides using a ruler does this need?"*). Note that the target's comment on the painting was not positive in both conditions, thus, different responses could be attributed to the offensive language in the norm-violating condition and not to mere positivity of the target's response.

Participants could choose to write a message responding to the comment or else to “Skip and exclude” the author of the comment. Choosing “Skip and exclude” would exclude the author of the comment from all further interaction with the participant and was classified as *ostracism*. If participants responded to the comment, they were further asked to select the emoji that best captured what they wanted to communicate, out of seven options (see Table 1). Emojis were classified either a *confrontation* (negative emoji) or an *inclusion* (neutral or positive emoji), see Table 1 for how the different emojis were coded. Additionally, we used Claude 3.5 Sonnet to classify the written responses on whether they actively confronted the recipient with their norm-violation; the results from the classification will henceforth be referred to as the “text-coded” responses.

Table 1

Emojis that the participants could assign to describe their message with. The selected emoji determined whether the response was categorized as inclusion or confrontation.

Emoji	Description	Classification
	Like	Inclusion
	Love	Inclusion
	Agree	Inclusion
	Okay	Inclusion
	Sad	Confrontation
	Angry	Confrontation
	Dislike	Confrontation

Afterwards, as a manipulation check, participants were asked to rate how “friendly”, “rude”, “out of line”, and “improper” they thought the target person’s behavior was, each on a 7-point Likert scale from 1 = “not at all” to 7 = “very much”.

2.2. Results

2.2.1. Manipulation Check

Two t-tests showed that the target’s behavior in the norm-violating condition was rated both as more improper and more out of line than the target’s behavior in the norm-conforming condition, $t(152.03) = -15.06$, $p < .001$, ($M_{\text{norm-violating}} = 6.1$, $SD = 1.7$ vs. , $M_{\text{norm-conforming}} = 2.1$, $SD = 1.6$, $d = -2.4$), and $t(157.46) = -16.61$, $p < .001$, ($M_{\text{norm-violating}} = 6.0$, $SD = 1.7$ vs. , $M_{\text{norm-conforming}} = 1.9$, $SD = 1.4$, $d = -2.6$). Thus, the manipulation of the norm-violation of the target can be considered successful. Beyond these direct measures of perceived inappropriateness, the target’s behavior in the norm violating condition was also perceived as less “friendly”, $t(142.53) = 8.50$, $p < .001$, ($M_{\text{norm-violating}} = 1.6$, $SD = 1.4$ vs. , $M_{\text{norm-conforming}} = 3.7$, $SD = 1.6$, $d = 1.4$), and more “rude”, $t(130.78) = -11.32$, $p < .001$, ($M_{\text{norm-violating}} = 6.4$, $SD = 1.5$ vs. , $M_{\text{norm-conforming}} = 3.3$, $SD = 1.9$, $d = -1.9$), compared to the target’s behavior in the norm-conforming condition.

2.2.2. Responses towards norm-violating vs. norm-conforming targets

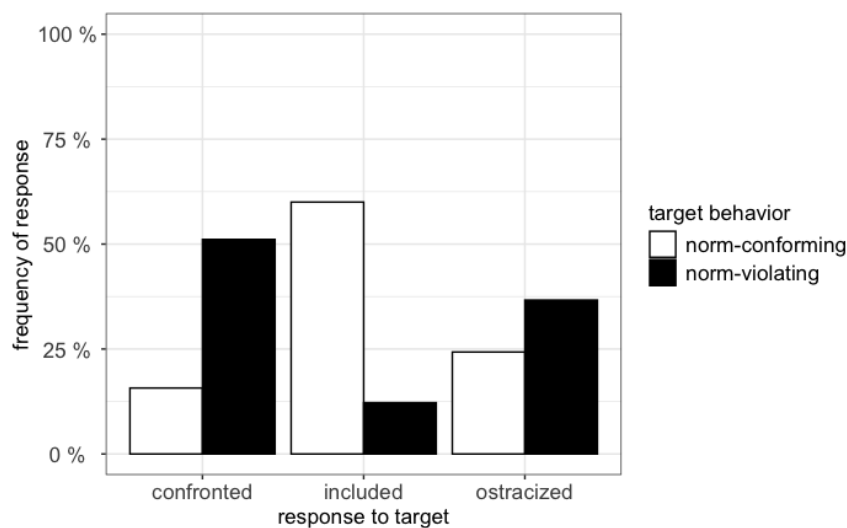
For the emoji-coded responses, the distribution of responses towards the norm-violating behavior was that 51.1% confronted the target, 36.7% ostracized the target, and 12.2% included the target. The distribution of the responses towards the norm-conforming behavior was that 15.7% confronted the target, 24.3% ostracized the target, and 60.0% included the target (see Figure 1).

A chi-square goodness of fit test indicated that responses towards the target were not distributed evenly across the conditions, $\chi^2(2) = 42.914$, $p < .001$, $w = 0.52$. To facilitate analyses, for the chi-square tests, participants’ decisions were re-coded into binomial variables (ostracism vs. no ostracism and confrontation vs. no confrontation). Descriptively, participants ostracized the target more often when they were norm-violating vs. norm-conforming, but the difference was not

statistically significant, $\chi^2(1) = 2.26, p = .133, w = 0.12, OR = 1.8 (0.9, 3.6)$. However, as hypothesized, there was a significant association between the norm-violating behavior of the target and whether or not they were confronted, $\chi^2(1) = 25.36, p < .001, w = 0.40$. The odds of confronting the target were 5.6 (2.6, 12.1) times higher if they showed norm-violating behavior than if they showed norm-consistent behavior. The frequencies of the response decisions towards the target are depicted in Figure 1.

Figure 1

Frequencies of the responses towards the target divided by target behavior (norm-violating vs. norm-conforming), in Study 1.



Note. The percentages represent the share of responses within each condition. In other words, each color adds up to 100%. The graph depicts the emoji-coded responses.

2.3. Within-subject Comparisons

As a less conservative test of whether norm-violating targets would be more likely to be ostracized than norm-conforming targets, we tested whether the target was ostracized significantly more often compared to the control person, depending on the experimental condition. In the norm-violating condition, the target was ostracized significantly more often, compared to the control person, McNemar's Chi-Squared: $\chi^2(1) = 12.97, p < .001, w = 0.28, OR = 4.2 (1.9, 8.9)$. This was not the case in the norm-conforming target condition ($\chi^2(1) = 1.90, p = .169, w = 0.11, OR = 1.9 (0.8, 4.6)$). We also

tested whether the target was confronted significantly more often compared to the control person, depending on the experimental condition. In the norm-violating condition, the target was confronted significantly more often, compared to the control person (McNemar's Chi-Squared: $\chi^2(1) = 36.21$, $p < .001$, $w = 0.48$, OR = 26.5 (7.78, 90.21). In the control condition, we employed Fisher's exact test (Warner, 2013), since no participants had confronted the control person, and the conditions to conduct a Chi-Square-Test were not met. The norm-conforming target was confronted significantly more often, compared to the control person (OR = Inf. (2.4, Inf.), $p < .001$, Cramer's V = 0.28).

2.4. Discussion

Study 1 showed that, as predicted, individuals choose to confront targets that violated an interaction norm significantly more often than targets that conformed to social norms. Although descriptively, the rate of ostracism was higher for norm-violating targets than for norm-conforming targets, the difference was not statistically significant. Moreover, exploratory analyses showed that the norm-violating target was ostracized significantly more often than the friendly control person. This indicates that the latent threshold for ostracism is a continuum on which few decide to ostracize a friendly person, while more decide to ostracize a negative and criticizing person and even more decide to ostracize a norm-violating person. In other words, in this paradigm, ostracism appears to not only be attractive when a person violates norms, but also when they are merely an undesirable interaction partner. This agrees with previous research, showing that people low in trait-agreeableness are undesirable interaction partners and thus more likely to be ostracized (Rudert et al., 2020, 2021).

Study 1 provided a first insight into the rate of confronting and ostracizing behavior following both norm-consistent and norm-violating behavior. While there was a significant increase in the rate of confrontation when the target was norm-violating, this was not the case for the rate of ostracism. The reason might be twofold: first, the target's behavior in the norm-conforming condition was perceived as unfriendly. Although this was not considered a norm-violation, it nonetheless made the target an undesirable interaction partner, as many participants indicated. Consequently, ostracism was also attractive in the norm-conforming condition. Second, mutually exclusive reactions of ostracism

and confrontation led to rather smaller effect sizes for analyses of the dichotomized variables, and thus required greater sample sizes. Thus, several changes were made in Study 2, to address these issues. Furthermore, we could not yet be sure whether participants interpreted emojis as we did, classifying “sad”, “angry”, and “dislike” as communicating negative emotion. We also addressed this in Study 2.

3. Study 2

Study 2 had a similar design to Study 1, with the following changes: To address the target’s negativity in the *norm-conforming* condition, we adapted the target’s comment in that condition to appear more friendly. Also, we increased the sample size to enhance the power of the statistical tests and enable the detection of smaller effect sizes.

3.1. Method

3.1.1 Participants and Design

All participants were recruited from Prolific Academic (US Americans only) to participate in a study on “Interaction online”, for a payment of GBP 1.25. Sample size was determined with GPower (Faul et al, 2007), according to which 263 participants provided sufficient power ($=.90$) to show a significant small effect ($w = 0.2$, $df = 1$) of norm-violating behavior on behavior choice. To account for potential data exclusions, we collected data until 280 participants had completed the study. One dataset was incomplete, and 4 persons indicated that their data should not be used for analysis, thus, the final sample consisted of 275 (196 female, 5 non-binary, $M_{\text{age}} = 34.6$, $SD = 13.2$) who were randomly assigned to two levels of a one-factorial between subject design (target behavior: norm-violating vs. norm-consistent).

3.1.2. Materials and Procedure

Materials and procedure were identical to Study 1, with the following changes: In the *norm-conforming* condition, the target’s comment was adapted to be more friendly, but still critical of the

painting ("I'm open to other opinions, but I struggle to recognize why much skill would be needed here. The motive seems kind of basic and the craftsmanship appears to be rather simple, when using the right tools, doesn't it?"). Finally, participants were asked to indicate the emotions they would try to convey, when using each of the emotions on a discrete scale containing, 1 = *negative*, 2 = *neutral*, or 3 = *positive*, to determine whether participants interpreted the emojis as we did, to validate our classification.

3.2. Results

3.2.1 Manipulation Check

Two t-tests showed that the target's behavior in the norm-violating condition was rated both as more improper and more out of line than the target's behavior in the norm-conforming condition, $t(270.82) = 18.13, p < .001, (M_{\text{norm-violating}} = 6.9, SD = 1.8 \text{ vs. } , M_{\text{norm-conforming}} = 2.3, SD = 1.6, d = 2.2)$ and $(t(255.59) = 16.88, p < .001, (M_{\text{norm-violating}} = 5.7, SD = 2.0 \text{ vs. } , M_{\text{norm-conforming}} = 2.2, SD = 1.5, d = 2.0) .$ Thus, the manipulation of the norm-violation of the target can be considered successful. Beyond these direct measures of perceived inappropriateness, the target's behavior in the norm violating condition was also perceived as less "friendly" , $t(272.10) = -12.75, p < .001, (M_{\text{norm-violating}} = 1.7, SD = 1.5 \text{ vs. } , M_{\text{norm-conforming}} = 4.0, SD = 1.5, d = -1.5)$, and more "rude", $t(272.51) = 14.60, p < .001, (M_{\text{norm-violating}} = 6.1, SD = 1.7 \text{ vs. } , M_{\text{norm-conforming}} = 3.1, SD = 1.8, d = 1.8)$, compared to the target's behavior in the norm-conforming condition.

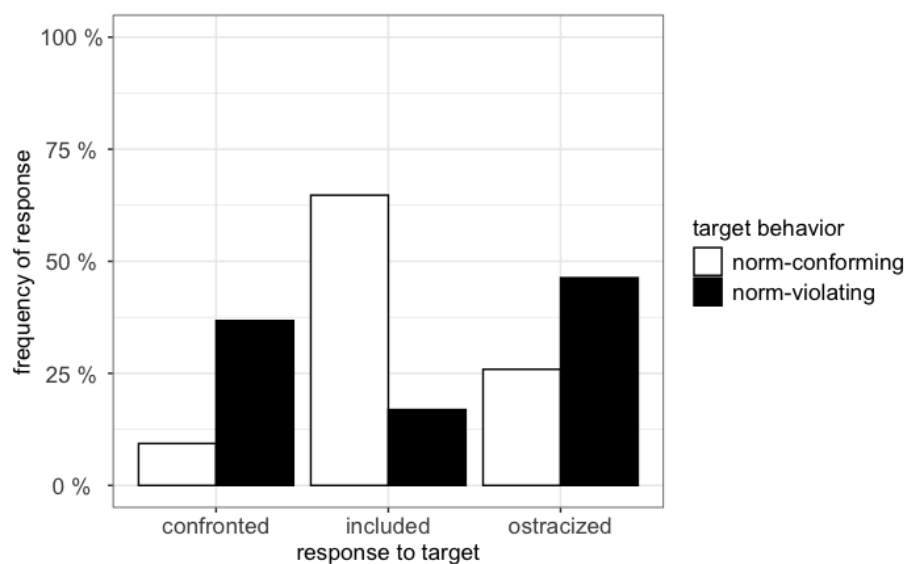
3.2.2. Responses towards norm-violating vs. norm-conforming targets

A Chi-square goodness of fit test indicated that response decisions towards the target were not distributed evenly across the conditions, $\chi^2(2) = 68.80, p < .001, w = 0.50$. To facilitate analyses, for the chi-square analyses, choices were again recoded into binomial variables (ostracism: yes vs. no; confrontation: yes vs. no).

In accordance with our hypotheses, there was a significant association between behavior of the target and whether they were ostracized, $\chi^2(1) = 11.58, p < .001, w = 0.21$. The odds of ostracizing the target were 2.4 (1.5, 4.1) times higher if the target showed norm-violating behavior compared to norm-consistent behavior. Also, in accordance with hypotheses, there was a significant association between the behavior of the target and whether they were confronted, $\chi^2(1) = 27.72, p < .001, w = 0.32$. The odds of confronting the target were 5.6 (2.9, 11.0) times higher if the target showed norm-violating behavior compared to norm-conforming behavior. The frequencies of the response decisions towards the target are depicted in *Figure 2*.

Figure 2

Frequencies of the responses towards the target divided by target behavior (norm-violating vs. norm-conforming).



Note. The percentages represent the share of responses within each condition. In other words, each color adds up to 100%. The graph depicts the emoji-coded responses.

3.3. Paradigm Evaluation

To assess the internal validity of our DV, we used several measures. First, we categorized the written responses to the target as either “confrontation” of the target, or an “inclusion” of the target by two independent raters (Cohen’s- $\kappa = 0.86$). The provided definition for confrontation was “any expression of a negative reaction towards the target or their comment” (see Supplement – Coding

Strategy, p. 1). Here, a negative reaction was further specified as criticism or derogation of the target or their behavior or opinion – signaling disagreement was not a sufficient condition for confrontation, as the target’s opinion was considered as “accepted, unless stated otherwise” (see OSF, *Data Analysis - Text Coded Prompt*). Disagreements between the raters’ classifications were resolved after the initial coding. The Spearman-Brown-Coefficient between the emoji-based categorization and the message-based categorization showed good internal consistency ($\rho = 0.79$). In total, more participants selected emojis classified as confrontational ($n = 65$) than confronted the target via message ($n = 43$). This seems to indicate that the coded messages are a more conservative measure of confrontation, as the barrier to openly criticize the target seems to be higher than assigning a confrontational emoji to one’s message.

As a second test of the internal validity of our dependent variable, we tested whether participants would indeed assign negative emotional valence to the emojis we had classified as confrontational and non-negative valences (i.e., positive or neutral) to the emojis we had classified as inclusionary. Thus, we asked participants which emotional valence (i.e., positive, neutral, negative) they would assign to each of the presented emojis. Indeed, most participants rated the emojis that were categorized as confronting as conveying negative emotion (sad = 97.5%, angry = 89.5%, dislike = 96.7%).

3.4. Discussion

The second study clarified some of the uncertainties that had remained after the first study. According to hypotheses, both ostracism and confrontation were shown more frequently towards norm-violating targets than towards norm-conforming targets. The obtained effect size of the target behavior – ostracism link ($w = 0.21$, OR (95%-CI) = 2.4 (1.5, 4.1)) further supports our assumption that the non-significant results in the first study were likely partly due to an underpowered test. Furthermore, our paradigm showed good internal validity based upon two assessments. First, the internal consistency between the dependent variable, based on the emoji that one assigned to their message, and the content of the messages, that were coded according to the construct definitions,

was high. Second, most participants rated the emojis that were classified as *confrontational*, as communicating negative emotions, while all other emojis were rated as communicating neutral or positive emotions. In summary, we developed a paradigm that allows us to investigate ostracism and confrontation decisions at the same time, as responses to norm-violating behavior and furthermore showed that both responses are attractive options to deal with a norm-violating target.

4. Study 3

In the third study we aimed to investigate whether the responses to norm-violating behavior in the online forum would be influenced by the CoC. We compared the effects of a *conflict-discouraging* CoC to the effects of a *conflict-encouraging* conduct, and no explicit CoC (control). We predicted that the rate of ostracism would *increase* with a prevailing conflict-discouraging CoC, relative to a prevailing conflict-encouraging CoC (Hypothesis 1) and no explicit CoC (Hypothesis 2). Additionally, we predicted that confrontation would *increase* with a prevailing conflict-encouraging CoC, relative to a prevailing conflict-discouraging CoC (Hypothesis 3) and *decrease* when there is a prevailing conflict-discouraging vs. no explicit CoC (Hypothesis 4). Taken together, we thus assumed that rather than contributing to a more friendly and conflict-free climate, conflict-discouraging CoCs might foster ostracism as an alternative means of norm enforcement. However, as a competing hypothesis, we further tested the possibility that inclusion might also *increase* with a prevailing conflict-discouraging CoC relative to a prevailing conflict-encouraging CoC (Hypothesis 5) and no explicit CoC (Hypothesis 6).

4.1. Method

4.1.1. Participants and Design

All participants were recruited from Prolific Academic (US Americans only) to participate in a study on “Interaction online”, for a payment of £1.25. Sample size was determined with a Monte-Carlo simulation (Muthén & Muthén, 2002) ($n = 5000$), according to which 150 participants per group (total $N = 450$) provided sufficient power ($=.82$) to show the smallest assumed effect size ($w = 0.16$, $df = 1$) of

norm-violating behavior on behavior choice. To account for potential data exclusions, we collected data until 500 participants had completed the study. 13 persons indicated that their data should not be used for analysis, thus, the final sample consisted of 487 US-Americans (240 female, 12 non-binary, $M_{age} = 36.8$, $SD = 13.4$) who were randomly assigned to three levels of a one-factorial between subject design (CoC: *conflict-discouraging* vs. *conflict-encouraging* vs. *none*).

4.1.2. Materials and Procedure

The materials were similar to Study 2, but with some crucial changes. To investigate the effects of CoCs on responses to norm-violations, participants in the two norm conditions now received an explicit instruction about how they should interact with the other participants: In the *conflict-discouraging* CoC condition, participants were told that “Every opinion is valid and should be accepted”, and that “We want the interaction in the art gallery to be positive.” In contrast, in the *conflict-encouraging* CoC condition, participants were told that “Every opinion is subjective and should be challenged”, and that “We want the interaction in the art gallery to be stimulating.” In the control condition, no explicit CoC was stated. The full instructions can be found in *Table 2*. The CoC instruction was presented after participants had supposedly been connected to the other participants and before they could comment on their experience with a painting. To reinforce the norm instruction, participants were asked to repeat the content of the presented CoC in one sentence, if a CoC was presented. Additionally, as a manipulation check, participants were asked to rate the following three statements: “In the art gallery, it was explicitly expected that other’s opinions would be [accepted/challenged].”, or “There were no explicit expectations about how to react to other’s opinions”, each on a 7-point Likert scale from 1 = “not at all”, to 7 = “very much”. The CoC was the only experimental factor, as we no longer manipulated whether the target was norm-violating or norm-conforming.

Table 2

Code of Conduct in every condition.

Condition	Code of Conduct
-----------	-----------------

Conflict-discouraging	Every opinion is valid and should be accepted! We want the interaction in the art gallery to be positive. Keep in mind that opinions that you do not agree with are always valid and you should accept them. We believe that there is benefit in respecting other opinions that you do not agree with, so you should accept those opinions.
Conflict-encouraging	Every opinion is subjective and should be challenged! We want the interaction in the art gallery to be stimulating. Keep in mind that opinions that you do not agree with are always subjective and you should challenge them. We believe that there is benefit in testing other opinions that you do not agree with, so you should challenge those opinions.
Control	<i>No explicit CoC</i>

Furthermore, the order of the (norm-violating) target person and the (norm-conforming) control person was randomized, to ensure that observed effects would not depend on the order of presentation.

4.2. Results

4.2.1. Manipulation Check

As intended, participants in the conflict-discouraging COC condition agreed with being told to accept other opinions more than in the other two conditions ($t(484) = 20.50, p < .001, d = 1.98$). In the conflict-encouraging condition, participants agreed more with being told to challenge other opinions than in the other two conditions ($t(484) = 23.43, p < .001, d = 2.24$) and in the control condition (no explicit norm), participants agreed more with not being told explicitly about how to interact, compared to the other two conditions ($t(484) = 16.33, p < .001, d = 1.57$).

4.2.2. Responses towards norm-violating vs. norm-conforming targets

To facilitate analyses, for the chi-square analyses, choices were again recoded into binomial variables (ostracism: yes vs. no; confrontation: yes vs. no; inclusion: yes vs. no). Below, all hypothesis tests are summarized in *Table 3* and all responses divided by condition are depicted in *Figure 3*.

4.2.2.1. Ostracism

A Chi-square goodness of fit test indicated that the rate of target ostracism was not distributed evenly across the conditions, $\chi^2(2) = 35.20$ $p < .001$, $w = 0.27$. In line with Hypothesis 1, stating that the rate of ostracism would *increase* with a conflict-discouraging CoC, (vs. a conflict-encouraging CoC), there was significantly more ostracism in the conflict-discouraging condition compared to the conflict-encouraging condition, OR = 3.7 (2.3, 5.9). Hypothesis 2, stating that the rate of ostracism would *increase* with a conflict-discouraging CoC (vs. no explicit norm), was not supported by the data: There was not significantly more ostracism in the conflict-discouraging condition compared to the control condition, OR = 1.1 (0.7, 1.8). An exploratory test, contrasting the rate of ostracism in the conflict-encouraging and the control condition with no norm showed that ostracism was OR = 3.2 (2.0, 5.2) times more likely in the control condition, $\chi^2(1) = 23.34$, $p < .001$, $w = 0.27$.

4.2.2.2. Confrontation

A Chi-square goodness of fit test indicated that the rate of target confrontation was not distributed evenly across the conditions, $\chi^2(2) = 21.28$ $p < .001$, $w = 0.21$. In accordance with Hypothesis 3, there was significantly more confrontation in the conflict-encouraging condition compared to the conflict-discouraging condition. This holds for both the frequencies in the emoji-coded responses (3a), OR = 2.9 (1.8, 4.6), and the text-coded responses (3b), OR = 3.4 (1.9, 6.0). Hypothesis 4, was not supported, as there was not significantly less confrontation in the conflict-discouraging condition compared to when there was no explicit norm. Again, this held for both the frequencies in the emoji coded responses (4a), OR = 0.7 (0.4, 1.2), and the text-coded responses (4b), OR = 0.6 (0.3, 1.1). An exploratory test, contrasting the rate of ostracism when the norm was conflict-encouraging vs. no norm showed that ostracism was OR = 3.2 (2.0, 5.2) times more likely in the control condition, $\chi^2(1) = 23.34$, $p < .001$, $w = 0.27$.

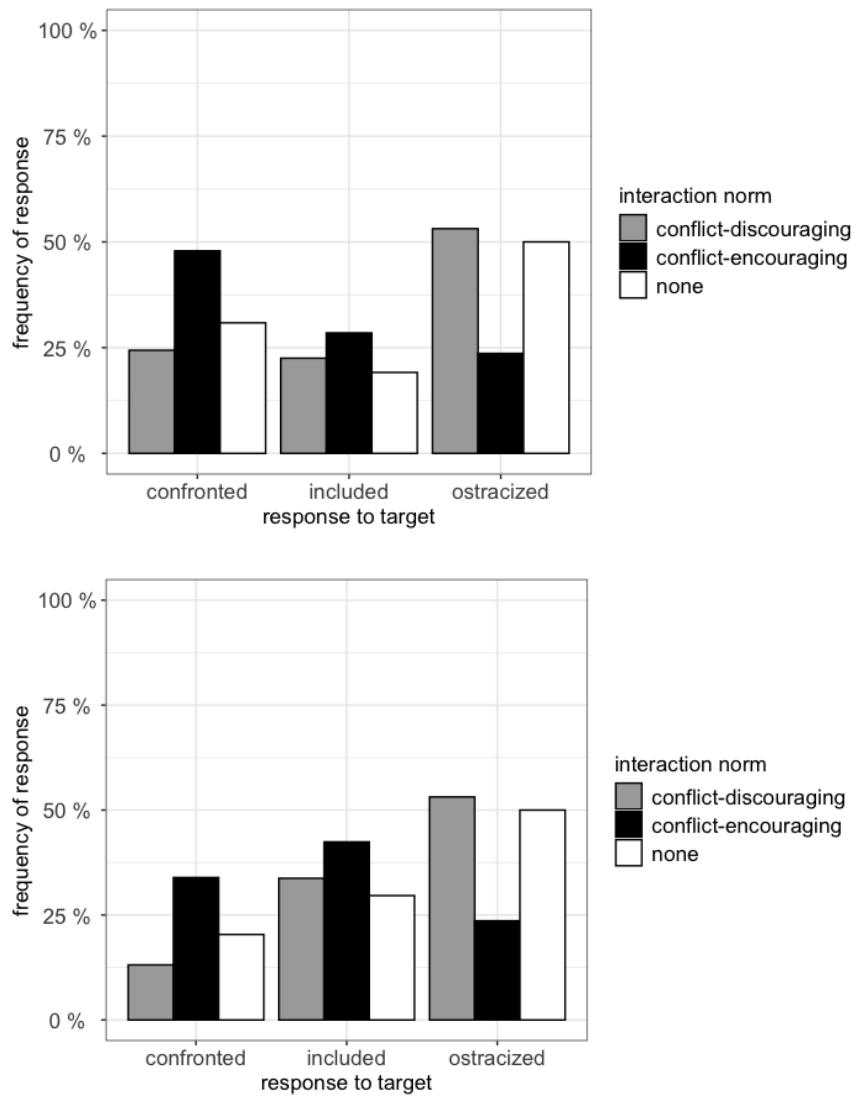
4.2.2.3. Inclusion

For *emoji-coded* responses, a Chi-square goodness of fit gave no indication that inclusion was unevenly distributed among the conditions, $\chi^2(2) = 4.09$ $p = .129$, $w = 0.09$. Equivalence tests (Wellek, 2002), further showed that there is no significant increase in inclusion rate when there was a conflict-discouraging norm. Following Wellek's (2002) procedure for non-equivalence tests for dichotomous outcomes, the 95% log-OR-CI for conflict-discouraging vs. conflict-encouraging norms, $(-0.82, 0.19)$, did not include the liberal bounds of 0.82, or strict bounds of 0.41, thereby strongly indicating that we can assume that the rate of inclusion, as indicated by non-negative emojis, is not greater when there is a conflict-discouraging CoC vs. a conflict-encouraging CoC. The 95% log-OR for conflict-discouraging CoC vs. no explicit CoC $(-0.33, 0.74)$ did not include the liberal bounds of 0.82, but did include the strict bounds of 0.41, thereby weakly indicating that we can assume that the rate of inclusion, as indicated by non-negative emojis, is not greater when there is a conflict-discouraging CoC vs. no explicit CoC.

For *text-coded* responses, a Chi-square goodness of fit test indicated that the rate of target inclusions was not distributed evenly across the conditions, $\chi^2(2) = 6.11$ $p = .047$, $w = 0.11$. Not conforming with our hypotheses, the significant difference came neither from the contrast between the conflict-discouraging vs. conflict-encouraging norm, $\chi^2(1) = 2.23$, $p = .135$, $w = 0.08$, nor the contrast between the conflict-discouraging vs. no norm, $\chi^2(1) = 0.46$, $p = .500$, $w = 0.04$. Rather, there was a significantly higher rate of inclusion when the norm was conflict-encouraging vs. when there was no norm, $\chi^2(1) = 5.26$, $p = .022$, $w = 0.13$. For an overview of the responses, separated by norm, see Figure 3.

Figure 3

Frequencies of the responses towards the target divided by the CoC, for emoji-coded confrontation (top) and text-coded confrontation (bottom).



Note. The percentages represent the share of responses within each condition. In other words, each color adds up to 100%.

Table 3

Significance tests and effect sizes for all Hypotheses in Study 3. CD = conflict-discouraging and CE = conflict-encouraging.

Hypothesis	Chi-squared test	OR (95%-CI)
1 - Ostracism for CD > CE	$\chi^2(1) = 28.70, p < .001, w = 0.30$	3.7 (2.3, 5.9)
2 - Ostracism for CD > None	$\chi^2(1) = 0.20, p = .653, w = 0.03$	1.1 (0.7, 1.8)
3a - Confrontation for CE > CD	$\chi^2(1) = 18.40, p < .001, w = 0.24$	2.9 (1.8, 4.6)

3b - Confrontation for CE > CD	$\chi^2(1) = 18.33, p < .001, w = 0.24$	3.4 (1.9, 6.0)
4a - Confrontation for CD < None	$\chi^2(1) = 1.39, p = .239, w = 0.07$	0.7 (0.4, 1.2)
4b - Confrontation for CD < None	$\chi^2(1) = 2.53, p = .112, w = 0.09$	0.6 (0.3, 1.1)
5a - Inclusion for CD > CE	$\chi^2(1) = 1.23, p = .267, w = 0.06$	0.7 (0.4, 1.2)
5b - Inclusion for CD > CE	$\chi^2(1) = 2.23, p = .135, w = 0.08$	0.7 (0.4, 1.1)
6a - Inclusion for CD > None	$\chi^2(1) = 0.37, p = .544, w = 0.03$	1.2 (0.7, 2.1)
6b - Inclusion for CD > None	$\chi^2(1) = 0.45, p = .500, w = 0.04$	1.2 (0.8, 1.9)

Note. Hypotheses 3a, 4a, 5a, and 6a are referring to the emoji-coded responses, while the corresponding b-Hypotheses are referring to the text-coded responses. Bold script indicates hypothesis-conform results. For all Odds Ratios, the Norm that is named first in the Hypothesis is in the numerator (e.g., CD > CE, OR = CD/CE)

4.3. Discussion

Study 3 provided evidence that CoCs influence how individuals respond to observed norm violations in online forums. More specifically, ostracism of a norm-violating target occurs more frequently when norms are conflict-discouraging and demand acceptance of others' opinions even in the case of disagreement. Confrontation of a norm-violating target on the other hand is more frequent when norms are conflict-encouraging and demand to challenge others' opinion in the case of disagreement. Interestingly, a conflict-discouraging norm did not foster inclusion of a norm-violating target compared to a conflict-encouraging norm.

Responses in the control condition, in which there was no explicit CoC, were more similar to the responses in the condition with a conflict-discouraging CoC than in the condition with a conflict-encouraging CoC. This suggests that, without further information, participants in online forums are cautious to confront, until it is explicitly communicated that this is accepted. But participants' behavior could always be linked to their Prolific-Account, which is not complete anonymity either, potentially making behavior more cautious. This seems to change when challenging others' opinions is explicitly encouraged and thereby socially beneficial. Therefore, in this specific online-setting, explicit conflict-

encouraging CoCs rather seemed to *prevent* ostracism, as opposed to explicit conflict-discouraging CoCs fostering it.

In sum, the third study showed that CoCs influence how we respond when we observe norm violations. When norms are conflict-discouraging, ostracism is more frequent and confrontation is less frequent, in comparison to when norms are conflict-encouraging.

5. Study 4

Studies 1-3 investigated responses to violations of interaction norms, that is, norms that aim to simplify interactions by providing structure and predictability. In the fourth study, we aimed to test the generalizability of the findings from Study 3 to a slightly different kind of norm violation, namely *generic exchange norms* - meaning norms that aim to regulate social exchange and cooperation (see Kerr & Levine, 2008). We therefore had to change the paradigm, as our current setting was designed primarily for inducing the violation of an interaction norm. As we wanted to replicate and generalize the findings from Study 3, our Hypotheses were generally similar, with minor adjustments in accordance with the results from Study 3: As in Study 3, we predicted that the rate of ostracism would *increase* with a conflict-discouraging CoC, relative to a prevailing conflict-encouraging CoC (Hypothesis 1). However, based on the findings of Study 3, we adapted Hypothesis 2, which now stated that ostracism would be less likely when the CoC would be conflict-encouraging (vs. none). Additionally, we again predicted that the rate of confrontation would *increase* with a conflict-encouraging CoC, relative to when there is a conflict-discouraging CoC (Hypothesis 3) and no explicit CoC (Hypothesis 4). Furthermore, in accordance with observations from Study 3, we now predicted that there would be *no effect* of the CoC on the rate of inclusion (Hypotheses 5&6), aiming at a replication of the equivalence tests from Study 3.

5.1. Method

5.1.1. Participants and Design

Participants were recruited from Prolific Academic (US Americans only) to take part in a study on “Interaction online”, for a payment of GBP 1.00. Sample size was determined with GPower (Faul et al, 2007), according to which 396 participants (132 per group) provided sufficient power ($=.90$) to show our smallest hypothesized effect, as assumed from the effects of Study 3, of the CoC on target confrontation ($w = 0.21$, $df = 2$). To account for potential data exclusions, we collected data from 424 participants. Eleven persons indicated that their data should not be used for analysis, thus, the final sample consisted of 413 US-Americans (221 female, 10 non-binary, $M_{age} = 40.6$, $SD = 13.9$) who were randomly assigned to three levels of a one-factorial between subject design (CoC: conflict-discouraging vs. conflict-encouraging vs. none).

5.1.2. Materials and Procedure

As typical online forums center around social interaction, and rarely contain mechanisms for (material) exchange, we decided to present participants with an example of an indirect norm-violation, to maximize ecological validity of the online forum setting. Particularly, rather than acting rude as in the *ArtEval* paradigm, the target reported an instance of having behaved highly uncooperative towards others in the past. However, while the topic of discussion was thus exchanged, the general setup of the task remained highly parallel to *ArtEval* from Studies 1-3. Specifically, the online forum now centered around the topic of how participants dealt with (or would deal with) shortages of different fundamental resources. After having been introduced to the task, participants were first asked whether they had experienced one of the shortages that would be the topic of discussion (e.g., “I have experienced a shortage of [medical supplies/ hygiene articles/...].”). In line with Studies 1-3, the experimental manipulation of the CoC was placed as part of the instruction after participants had supposedly been connected to the other participants and before they could comment on their experience with a shortage. Participants would then first be asked to describe how they acted, or would act, during a shortage of medical supplies. They were then told that they would now read other

participants' responses to other kinds of shortages. Again, all participants were exposed to two responses - one norm-conforming response and one norm-violating response, with the order of the two being randomized. The norm-conforming participant clearly conformed to - or arguably even exceeded - generic exchange norms, by reporting that, in case of a gasoline shortage, they would “only use [their] car when absolutely necessary”. The norm-violating participant on the other hand, reported: “I got pretty savvy during the toilet paper rush. Figured out which stores didn't enforce limits strictly and made the rounds regularly. Sure, I ended up with more than I needed, but it's better to be over-prepared than under, right? My basement was always well-stocked.” This behavior was designed to be a violation of “communal sharing norms”, which are norms “designating uncertain resources as common properties to be shared with other members” (Kameda et al., 2003). The mechanism of response (emoji-coded and text-coded) was the same as in the previous studies.

5.2. Results

As intended, participants in the conflict-discouraging condition agreed with being told to accept other opinions more than in the other two conditions ($t(410) = 20.50, p < .001, d = 2.15$). In the conflict-encouraging condition, participants agreed more with being told to challenge other opinions than in the other two conditions ($t(410) = 16.19, p < .001, d = 1.70$) and in the control condition (no explicit norm), participants agreed more with not being told explicitly about how to interact, compared to the other two conditions ($t(410) = 11.57, p < .001, d = 1.20$).

5.2.1. Responses towards norm-violating vs. norm-conforming targets

To facilitate analyses, for the chi-square analyses, choices were again re-coded into binomial variables (ostracism: yes vs. no; confrontation: yes vs. no; inclusion: yes vs. no). Below, all hypotheses test-results are summarized in *Table 4* and all responses divided by condition are depicted in *Figure 4*.

5.2.1.1. Ostracism

A Chi-square goodness of fit test indicated that the rate of target ostracism did not significantly differ between the conditions, $\chi^2(2) = 2.86$ $p = .239$, $w = 0.08$. Consequently, there was neither a significant increase in the rate of ostracism when the CoC was *conflict-discouraging* vs. *conflict-encouraging* (Hypothesis 1), OR = 1.4 (0.7, 2.6), nor did it decrease when it was *conflict-encouraging* vs. *no CoC* (Hypothesis 2), OR = 1.3 (0.6, 2.6).

5.2.1.2. Confrontation

A Chi-square goodness of fit test indicated that the rate of target confrontation did not significantly differ between the conditions, $\chi^2(2) = 3.92$, $p = .141$, $w = 0.10$. Consequently, there was neither a significant increase in the rate of confrontation when the CoC was *conflict-encouraging* vs. *conflict-discouraging* (Hypothesis 3a: emoji-coded), OR = 1.6 (0.9, 2.7), nor did it increase when it was *conflict-encouraging* vs. *no CoC* (Hypothesis 4a: emoji-coded), OR = 1.3 (0.6, 2.6). This pattern held for the text coded responses (see Table 4).

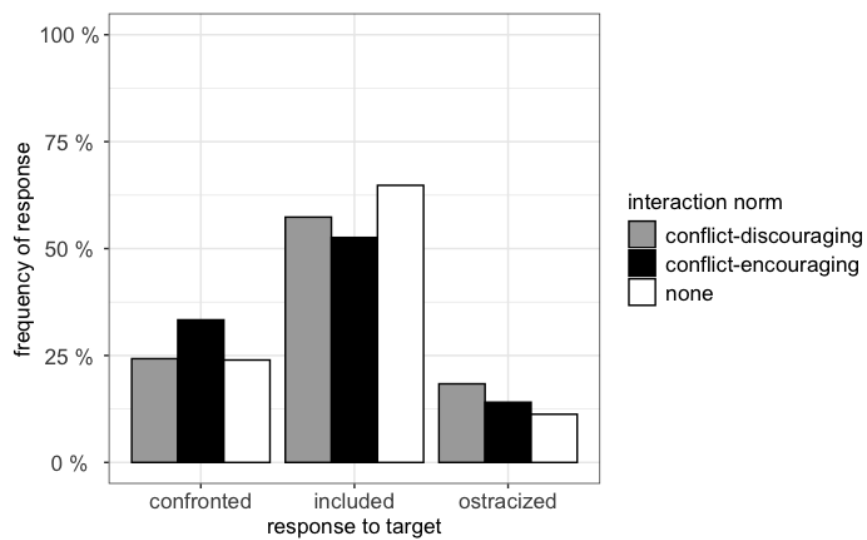
5.2.1.3. Inclusion

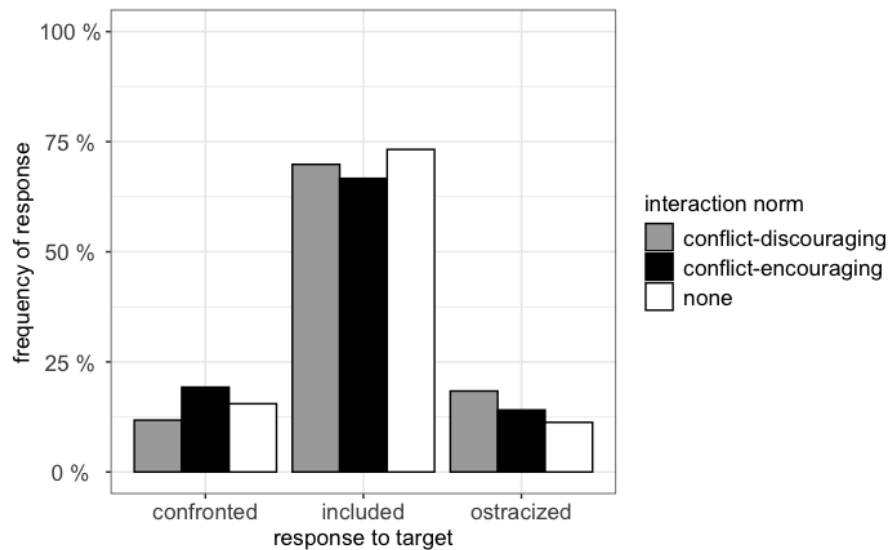
For *emoji-coded* responses equivalence tests showed that the differences in inclusion rates remained below the liberal equivalence criterion (0.82), for the conflict-discouraging vs. conflict-encouraging CoC (Hypothesis 5a), 95% log-OR-CI = (-0.29, 0.67) and below the strict equivalence criterion (0.41) for conflict-discouraging vs. no CoC (Hypothesis 6a), 95% log-OR-CI = (-0.79, 0.17). This pattern holds for the *text-coded* responses (see Table 4).

For *text-coded* responses, a Chi-square goodness of fit test indicated that the rate of target included was not distributed evenly across the conditions, $\chi^2(2) = 6.11$ $p = .047$, $w = 0.11$. Not conforming with our hypotheses, the significant difference came neither from the contrast between the conflict-discouraging vs. conflict-encouraging CoC, $\chi^2(1) = 2.23$, $p = .135$, $w = 0.08$, nor the contrast between the conflict-discouraging vs. no CoC, $\chi^2(1) = 0.46$, $p = .500$, $w = 0.04$. Rather, there was a significantly higher rate of inclusion when there was no CoC vs. when the CoC was conflict-encouraging, $\chi^2(1) = 5.26$, $p = .022$, $w = 0.13$. For an overview of the responses, separated by norm, see Figure 4.

Figure 4

Frequencies of the responses towards the target divided by the CoC, for the emoji-coded responses (top) and the text-coded responses (bottom).





Note. The percentages represent the share of responses within each condition. In other words, each color adds up to 100%.

Table 4

Significance tests and effect sizes for all Hypotheses in Study 3. CD = conflict-discouraging and CE = conflict-encouraging.

Hypothesis	Chi-squared test	OR (95%-CI)
1 - Ostracism for CD > CE	$\chi^2(1) = 0.63, p = .425, w = 0.05$	1.4 (0.7, 2.6)
2 - Ostracism for CE < None	$\chi^2(1) = 0.27, p = .602, w = 0.03$	1.3 (0.6, 2.6)
3a - Confrontation for CE > CD	$\chi^2(1) = 2.29, p = .130, w = 0.09$	1.6 (0.9, 2.7)
3b - Confrontation for CE > CD	$\chi^2(1) = 2.36, p = .124, w = 0.09$	1.8 (0.9, 3.5)
4a - Confrontation for CE > None	$\chi^2(1) = 2.55, p = .110, w = 0.10$	1.6 (0.9, 2.7)
4b - Confrontation for CE > None	$\chi^2(1) = 0.45, p = .503, w = 0.04$	1.3 (0.7, 2.4)
Hypothesis	Equivalence-Test	OR (95%-CI)
5a - Inclusion for CD $\approx$ CE	95% log-OR-CI = (-0.29, 0.67)	1.2 (0.8, 2.0)
5b - Inclusion for CD $\approx$ CE	95% log-OR-CI = (-0.36, 0.66)	1.2 (0.7, 1.9)
6a - Inclusion for CD $\approx$ None	95% log-OR-CI = (-0.79, 0.17)	0.7 (0.5, 1.2)

6b - Inclusion for CD $\approx$ None

95% log-OR-CI = (-0.69, 0.35)

0.9 (0.5, 1.4)

Note. Hypotheses 3a, 4a, 5a, and 6a are referring to the emoji-coded responses, while the corresponding b-Hypotheses are referring to the text-coded responses. Emphasized lines indicate hypothesis-conform results. For the equivalence test, 0.41 is the strict boundary and 0.82 is the liberal boundary for equality (Wellek, 2002).

5.3 Exploratory Analyses

Participants were a lot more divided about whether the target behaved out of line than they were for the previous studies - $M = 3.94$, $SD = 2.17$ (compared to Study 3: $M = 5.79$, $SD = 1.99$ - both on a scale from 1 to 7). Thus, we reanalyzed Hypotheses 1 and 3, while only including the participants that rated the targets behavior to at least be somewhat out of line (≥ 5 , $n = 128$). The hypotheses were not supported under these conditions either; ostracism - $\chi^2(1) = 0.76$, $p = .383$, $w = 0.08$, confrontation (emoji-coded) - $\chi^2(1) = 1.26$, $p = .262$, $w = 0.10$.

5.4. Discussion

The aim of Study 4 was to generalize the findings for responses to violations of interaction norms from Study 3 to a violation of generic exchange norms in Study 4. However, results of Study 4 neither suggested a significant impact of the CoC on the rate of ostracism, nor a significant impact of the CoC on the rate of confrontation. As hypothesized, we found evidence for an equivalence across the rates of inclusion over the different CoC. Yet, the results of the equivalence test are considerably less interesting, given that the CoCs generally had little impact on the responses to the observed norm-violation.

Generally, the rate of ostracism was descriptively much lower for this violation of a generic exchange norm than for the previous interaction norm-violation (see *Integrative Analysis* for more details), which was mirrored by a much higher rate of inclusion. One plausible explanation for this could be that a much larger share of participants did not consider the norm-violation to be “out of line”. Yet, even when filtering for participants that did find the behavior out of line, the rate of ostracism remained relatively low (16.4%). Such a lower base-rate makes it more difficult to appropriately power the statistical tests. Also, this begs the question whether this different rate in ostracism might be a general difference in the response tendencies towards the violation of an interaction norm vs. a generic exchange norm, or whether it is simply due to specificity of the paradigm. Clearing up this uncertainty, as well as replicating the findings from Study 3, were the main goals for the final study, Study 5.

6. Study 5

Study 5 aimed to 1) replicate the findings from Study 3 and 2) investigate differences in the responses to the violation of interaction norms and generic exchange norms. Consequently, we manipulated both factors experimentally. As this demanded to keep everything but the kind of norm-violation parallel, we needed a paradigm that allowed for both kinds of norm-violation, as well as the manipulation of the CoC. We decided to add a layer to the original paradigm from Studies 1-3, by

introducing a financial reward system that was then exploited by the target in the generic exchange norm-violation. As we now had a two-factor design, we decided to drop the control condition without a CoC, as we wanted to maximize the power for remaining conditions. Our hypotheses on the influence of the CoC on the responses remained the same as in Studies 3 and 4: We expected the rate of ostracism to increase when CoCs are conflict-discouraging (vs. conflict-encouraging - Hypothesis 1), and we expected the rate of confrontation to increase when the CoC is conflict-encouraging (vs. conflict-discouraging - Hypothesis 2). Finally, we had competing hypotheses on the generalizability of the effect of the injunctive interaction norms on the rate of ostracism (Hypothesis 3), stating that a) the effect of the injunctive interaction norms on ostracism will be stronger if the target has violated an interaction norm (vs. a generic exchange norm), or b) the effect will not be moderated by the type of norm-violation.

6.1. Method

6.1.1. Participants and Design

Participants were recruited from Prolific Academic (US Americans only) to take part in a study on “Interaction online”, for a payment of GBP 1.44. Sample size was determined with GPower (Faul et al, 2007), according to which 380 participants (190 per group) provided sufficient power ($=.90$) to show our smallest hypothesized effect, as assumed from the effects of Study 3, of the CoC (conflict-discouraging vs. conflict-encouraging) on target confrontation ($w = 0.24$, $df = 1$). We thus collected data from 381 participants, of which five persons indicated that their data should not be used for analysis. Thus, the final sample consisted of 376 US-Americans (204 female, 5 non-binary, $M_{age} = 38.64$, $SD = 12.30$) who were randomly assigned to four levels of a two-factorial between subject design (CoC: conflict-discouraging vs. conflict-encouraging x violated norm: interaction norm vs. generic exchange norm).

6.1.2. Materials and Procedure

As one of the main goals was to replicate our findings from Study 3, we built on the *ArtEval*-paradigm from Studies 1-3, integrating a layer that allowed for the manipulation of the kind of norm-violation. Specifically, targets in the generic exchange norm condition stole intellectual property by copying a comment from an alleged other participant. To make this more plausible, participants were told that independent jurors would rate all the original comments on the paintings, considering also how much engagement those comments created, and that the best comment would receive a “small bonus payment”. To allow for the possibility to copy a comment from another participant, the instructions further stated that it was possible to start a new comment thread for any particular painting, after having been introduced to the task, the task went on as before up to the point that the *kind of norm-violation* was manipulated. After having read the second comment on a picture, they were told that the target had decided to start another comment thread on this painting. The target’s comment in this supposedly newly started thread then differed, depending on the experimental condition: In the condition *violated interaction norm*, the norm-violation was the same as in Studies 1-3, with the target being unnecessarily rude towards the artist (see *Study 1*). In the condition *violated generic-exchange norm*, the target copied the comment from the previous contributor. To ensure that this would not be perceived as a technical error, we added a couple of words before the copied content. Specifically, the comment read: “There's more to the painting than one might think. *To me, this painting seems to simplify the city into basic shapes and warm colors. I get the impression the artist might be inviting us to find beauty in urban simplicity. The soft hues and sharp shadows create what feels like an inviting yet abstract cityscape.. But am I being too optimistic?*”, with everything in italics representing the comment of the previous contributor. After responding to the norm-violation, participants were asked to rate the behavior of the other participant as in the previous studies, as well as conducting an attention check on the perception of the manipulated kind of norm-violation. The rest remained constant to the Studies 1-3. The entire paradigm can also be seen in the supplementary information on the OSF (see *Materials*).

6.2. Results

6.2.1 Manipulation Check

6.2.1.1. Code of Conduct

Participants in the conflict-discouraging CoC conditions agreed with being told to accept other opinions more than participants in the conflict-encouraging CoC conditions ($t(375) = 17.71, p < .001, d = 1.29$). In the conflict-encouraging CoC conditions, participants agreed more with being told to challenge other opinions than in the conflict-discouraging CoC conditions ($t(375) = 13.69, p < .001, d = 0.99$).

6.2.1.2. Kind of Norm-Violation

Of the participants in the *interaction norm* condition, 84% (158/188) reported they realized the target's comment was rude towards the artist, indicating that, in this condition, the situation was perceived as intended. Of the participants in the *generic-exchange norm* condition on the other hand, only 49% (93/188) reported to have realized that the target's comment was stolen from a previous comment. This indicates that many of the participants had not realized that there was a norm-violation. We considered this by running the analyses with and without the participants who failed to detect the manipulation.

6.2.2. Responses towards norm-violating vs. norm-conforming targets

To facilitate analyses, for the chi-square analyses, choices were again recoded into binomial variables (ostracism: yes vs. no; confrontation: yes vs. no; inclusion: yes vs. no). Below, all hypotheses test-results are summarized in *Table 5* and all responses divided by condition are depicted in *Figure 5*.

6.2.2.1. Ostracism

There was no significant overall increase in the rate of ostracism when the CoC was *conflict-discouraging* vs. *conflict-encouraging* (Hypothesis 1), $OR = 1.2 (0.8, 1.8)$. This did not differ, when the participants with a failed manipulation check were excluded from the sample, $OR = 1.0 (0.6, 1.7)$. In a logistic regression, there was also no significant interaction between the CoC and the kind of norm-violation (Hypothesis 3), $OR = 1.7 (0.7, 4.0)$ (Here the OR refers to the odds of target exclusion for the combination of conflict-encouraging CoC and social-exchange norm violation. In other words, there's

a (non-significant) 70.0% increase in the odds of target exclusion for this specific combination, compared to what we would predict based on the main effects alone). Lastly, the rate of ostracism was not significantly different when only considering the conditions where an interaction norm was violated ($n = 190$), $OR = 1.6$ (0.8, 2.9).

6.2.2.2. Confrontation

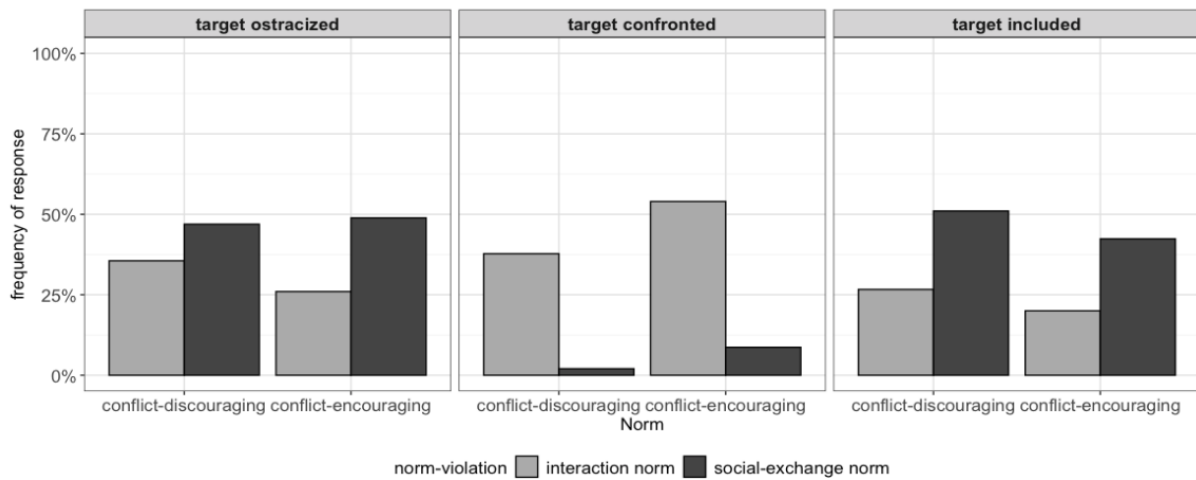
In accordance with Hypothesis 2, there was significantly more confrontation in the conflict-discouraging COC condition compared to the conflict-encouraging CoC condition. This holds for both the frequencies in the emoji-coded responses (2a), $OR = 2.0$ (1.3, 3.2), and the text-coded responses (2b), $OR = 2.1$ (1.2, 3.8). The effect sizes did not increase, when the participants with a failed manipulation check were excluded from the sample, emoji-coded: $OR = 1.7$ (1.0, 2.9); text-coded: 2.1 (1.1, 3.9).

6.2.2.3. Inclusion

For *emoji-coded* responses, there was no significant difference between the rate of inclusion in the conflict-discouraging condition vs. the conflict-encouraging condition, $OR = 1.5$ (1.0, 2.2), $\chi^2(1) = 2.74$, $p = .100$, $w = 0.09$. The effect size increased slightly, when the participants with a failed manipulation check were excluded from the sample, $OR = 1.7$ (1.0, 3.1), $\chi^2(1) = 3.79$, $p = .052$, $w = 0.09$, but the effect remained non-significant. The pattern was the same for the text-coded responses (see OSF for details).

Figure 5

Frequencies of the responses towards the target divided by the CoC and the kind of norm-violation.



Note. The percentages represent the share of responses within each of the 2x2 conditions. In other words, each color for each CoC adds up to 100%. The graph depicts the emoji-coded responses.

Table 5

Significance tests and effect sizes for all Hypotheses in Study 3. CD = conflict-discouraging and CE = conflict-encouraging.

Hypothesis	Chi-squared test	OR (95%-CI)
1 - Ostracism for CD > CE	$\chi^2(1) = 0.63, p = .426, w = 0.04$	1.2 (0.8, 1.8)
2a - Confrontation for CE > CD	$\chi^2(1) = 7.90, p = .004, w = 0.12$	2.0 (1.3, 3.2)
2b - Confrontation for CE > CD	$\chi^2(1) = 6.06, p = .014, w = 0.13$	2.1 (1.2, 3.8)
3 - H1 is dependent on <i>Kind of NV</i>	$b = 0.53, SE = 0.43, p = .217$	1.7 (0.7, 4.0)

Note. Hypotheses 2a is referring to the emoji-coded responses, while the corresponding b-Hypotheses are referring to the text-coded responses. Emphasized lines indicate hypothesis-conform results.

6.2.3. Was stealing the comment considered a norm-violation?

Was copying the comment considered a Norm-Violation? While participants in the violated *interaction norm* condition considered the behavior by the target to be rather out of line ($M = 5.54$, $SD = 2.00$, on a scale from 1 to 7), with the majority of people ($88/160 = 55\%$) reporting a maximum of 7, this was not the case for the participants in the participants in the violated generic-exchange norm

condition ($M = 1.61$, $SD = 1.35$), where 76% of participants reported the minimum of 1. Also, these numbers did not change much when we only considered participants that had at least reported to have recognized that the comment had been copied ($M = 1.61$, $SD = 1.39$, Reported 1: 73/94 = 78%).

6.3. Discussion

Study 5 answered some open questions satisfyingly, while posing some new ones. As for the main focus of this paper, the impact of CoCs on norm-enforcement behavior, the findings were mixed. The rate of ostracism was not significantly higher when the CoC was conflict-discouraging, as had been the case in Study 3. Yet, it would also be rash to conclude that the results in Study 3 were a fluke. When considering the interaction norm condition only (i.e. the condition that essentially replicated Study 3), the effect was smaller than before and thus failed to reach significance given the smaller sample size, yet the descriptive trend remained. The picture was a lot clearer for the rate of confrontation, with the rate of both emoji- as well as text-based confrontation being higher when CoCs were conflict-encouraging (vs. conflict-discouraging), as predicted by Hypothesis 2.

One limitation of Study 5 was many participants did not perceive copying a comment as violating any norms. Some participants signaled that the comment was stolen (e.g. "You literally copied and pasted the last person's response to this picture, which you commented on.", "This is the same thing [target] said"), and all participants that did report this reported much higher inappropriateness of the target (≥ 6 , in contrast with an overall mean of $M = 1.61$). They could be more sensitive to intellectual property theft, but, given how much their ratings deviate from the rest of the distribution, it seems much more likely that they were among the minority that actually realized the norm-violation. While we can only speculate on the origins of this, none of the interpretations invalidates the fact that this prohibits a proper comparison of the two norm-violation conditions.

While some overall trends were clearer than others, the results diverged somewhat between Studies 3-5, obscuring potential general trends, between less robust chance outcomes. To outline a more comprehensive overview of the findings from Studies 3-5, in the following we want to outline some of these overall trends more clearly, in an integrative analysis (Curran & Hussong, 2009).

7. Integrative Analysis

Here, we will consider the overall results, considering all the relevant data that can be informative to answer the major questions discussed in this paper. Such a Within-Paper-Meta Analysis will allow us to gain the maximal statistical power for our statistical analyses.

7.1 Codes of Conduct and Norm-Enforcement

We will consider the data from Studies 3-5, all studies where the CoC was manipulated. To keep the analysis to the point, we will focus on the contrasts between a conflict-discouraging and a conflict-encouraging CoC.

7.1.1 Ostracism

Overall ($N = 976$), the rate of ostracism was $OR = 1.8$ (1.4, 2.4) times higher when the CoCs were conflict-discouraging (vs. conflict-encouraging), $\chi^2(1) = 17.16$, $p < .001$, $w = 0.13$. We specified a logistic regression model, including the CoC, the kind of norm violation, and their interaction. Here, the rate of ostracism was $OR = 1.8$ (1.4, 2.5) times higher for interaction norm-violations than for violations of generic-exchange norms, $b = 0.70$, $p < .001$. Also, there was a significant interaction, $b = 0.89$, $p = .001$, according to which ostracism in the conflict-discouraging condition was $OR = 2.4$ (1.4, 4.2) times more likely when an interaction norm was violated, compared to when a generic-exchange norm was violated. This was reflected in the fact that the overall main-effect of the CoC was only significant for the violation of interaction norms ($n = 515$), $OR = 2.7$ (1.9, 3.9), $\chi^2(1) = 26.96$, $p < .001$, $w = 0.23$, but not for violations of generic-exchange norms ($n = 461$), $OR = 1.1$ (0.7, 1.7), $\chi^2(1) = 0.16$, $p = .686$, $w = 0.02$.

7.1.2. Confrontation

Overall, the rate of ostracism was $OR = 2.1$ (1.6, 2.8) times higher when CoCs were conflict-encouraging (vs. conflict-discouraging), $\chi^2(1) = 27.08$, $p < .001$, $w = 0.17$. We specified a logistic

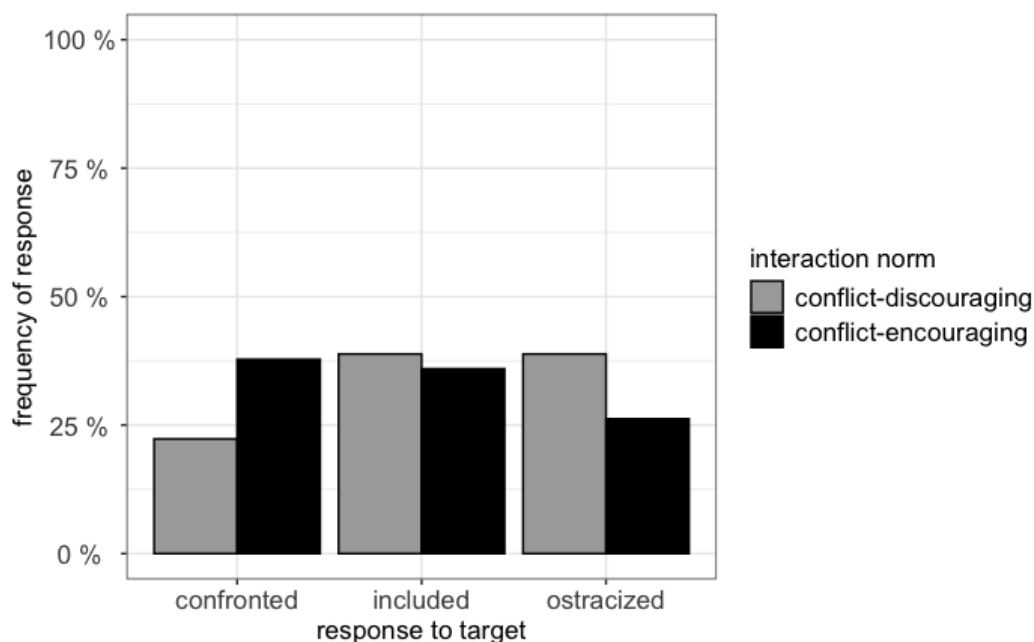
regression model, including the CoC, the kind of norm violation, and their interaction. Here, the rate of ostracism was $OR = 2.3$ (1.5, 3.7) times higher for interaction norm-violations than for violations of generic-exchange norms, $b = 0.85$, $p < .001$. There was no significant interaction though, $b = 0.34$, $p = .258$, which is reflected in the fact that the overall main-effect of the CoC was significant for both violations of interaction norms ($n = 515$), $OR = 2.4$ (1.7, 3.5), $\chi^2(1) = 22.75$, $p < .001$, $w = 0.21$, as well as for violations of generic-exchange norms ($n = 461$), $OR = 1.7$ (1.1, 2.8), $\chi^2(1) = , p = .030$, $w = 0.11$.

7.1.3. Inclusion

Overall, the rate of inclusion was not significantly higher in any of the two CoC conditions, $\chi^2(1) = 0.74$, $p = .390$, $w = 0.03$, $OR = 1.1$ (0.7, 1.6) (OR for a conflict-encouraging vs. conflict-discouraging CoC). This remained true for both violations of interaction norms, $\chi^2(1) = 0.06$, $p = .814$, $w = 0.01$, and generic-exchange norms, $\chi^2(1) = 1.55$, $p = .212$, $w = 0.06$. Yet, there was a significant impact of the kind of norm-violation on the rate of inclusion, $b = 1.34$, $p < .001$, as the rate of inclusion was $OR = 3.8$ (2.6, 5.7) times higher for the observed violations of generic-exchange norms than observed violations of interaction norms.

Figure 6

Frequencies of the responses towards the target divided by the interaction norm.



7.2. Discussion

The integrative analysis shows a general trend for an increase in the rate of ostracism when the CoC is conflict-discouraging, and an increase in confrontation when the CoC is conflict-encouraging. While the effect on the rate of confrontation held for both the observed violation of interaction norms and the observed violation of generic-exchange norms, the effect on the rate of ostracism was only observable when the violation of interaction norms was observed. While there was a main-effect of the kind of norm-violation on the rate of inclusion, this has to be interpreted with caution, as the many of the participants in the Study 5 had reported not having seen the norm-violation, and even many of those that reported so, did not consider it to be “out of line”, which was mostly the case for the violation of the interaction norm.

8. General Discussion

CoCs are ubiquitous in online forums, and while such often highly restrictive in their efforts to suppress unwanted contributions (e.g., Chandrasekharan et al., 2018), little is known about potential adverse consequences this can have on the social processes in the virtual environment. Here, we investigate how such CoCs affect norm-enforcement in online forums, specifically contrasting conflict-encouraging and conflict-discouraging CoCs (cf. Gibson, 2019), in simulated experimental settings. Particularly, we show that the rate of ostracism of norm-violating targets is higher when CoCs are conflict-discouraging - i.e., “Every opinion is valid and should be accepted!”, while rates of confrontation of norm-violating targets are higher when CoCs are conflict-encouraging - i.e., “Every opinion is subjective and should be challenged!”. The rate of inclusion was not significantly affected by the CoC, suggesting that conflict-discouraging CoCs, instead of fostering more positive interaction, rather lead people in online forums to decide not to interact at all. These effects were more stable when an *interaction norm* was violated (i.e. rude language, Studies 1-3, 5), compared to when a *generic-exchange norm* was violated (i.e., egoistic behavior, Studies 4-5). It should be noted though,

that egoistic behavior was identified and classified as a norm-violation much less reliably than the rude language, making it difficult to draw any robust conclusions.

8.1. Conflict-Discouraging Codes of Conduct Can Facilitate Ostracism

Previous literature on the influence of conflict-discouraging CoCs on behavior in online forums was limited to the mere presence or absence of activity (Gibson, 2019). Here, the pattern of engagement and disengagement generally matched the observations from Gibson (2019), where the conflict-discouraging and more restrictive “safe space” subreddit entailed less engagement in comparison to the more conflict-encouraging and less restrictive “free speech” subreddit. The experimental control in our studies allowed us to pin down one specific instance of disengagement when the CoC was conflict-discouraging: Deciding to ostracize a norm-violating interaction partner rather than confronting them. This result concurs with Archer and Coyne’s (2005) framing of ostracism as the subtle and less risky alternative to directly confronting somebody about their behavior. Here, the risk being that one might become a norm-violator oneself, as confrontation could be inappropriate in forums with conflict-discouraging CoCs (Eriksson et al., 2017; Strimling & Eriksson, 2014).

8.1.1. Is this an issue?

The finding that ostracism may become attractive when CoCs are more restrictive and conflict-discouraging, is not without cause for concern: First of all, what ostracism does not offer much in terms of explicit feedback and its positive consequences. In contrast, conflict is highly effective at changing opinions (e.g., Johnson & Johnson, 2009), even when eliciting negative emotions (Czopp et al., 2006). However, it could be argued that, particularly in online forums, algorithms reward a kind of confrontation that is characterized by moral outrage, making it unlikely to lead to constructive conflict (Brady et al., 2017; Brady & Crockett, 2024; Brady, Gantman, & Van Bavel, 2020), and that ostracism is also a proven tool to correct unwanted behavior (Hales et al., 2017). This is reflected in the often quoted internet-rule “*Don’t feed the trolls*” that essentially encourages users to not engage those that violate norms (and thus ostracize). But even without considering its effectiveness, confrontation is often more clear about what exactly is disapproved of. In comparison, ostracism requires the target to

correctly interpret why they were punished, and thus always comes with a high amount of ambiguity for the target ambiguity (Williams, 2009). This ambiguity and the uncertainty it can elicit are also the grounds for an argument against ostracism, from an emotional standpoint. People systematically underestimate the severity of the negative emotional consequences of ostracism (Nordgren et al., 2011; O'Reilly et al., 2015) and it has repeatedly been found that even negative attention is perceived as less aversive than not receiving any attention at all (O'Reilly et al., 2015; Rudert et al., 2017). While additional research is needed on this topic, we assume that confrontation could at least be more targeted so that the target knows why they have offended others, thus preventing the uncertainty and pain often related to experiences of ostracism.

8.2. The Power of Codes of Conduct

While the overall effect of ostracism was rather small ($OR = 1.8$ (1.4, 2.4), between a small OR ($=1.68$) and medium OR ($=3.47$)), that we were cautious to not overemphasize the CoCs in an unrealistic way to not threaten the ecological and external validity of the study. Specifically, participants were told about the CoC once at the beginning of the study and then once more before responding to the first comment (which was not the response to the target, in most cases). That the salience of the CoC could play a role on their impact has been supported by data from within and without the study. Our own data shows that the effects of the interaction norm on the norm-enforcement behavior were stronger in Study 3 compared to Study 5. The greatest difference between the studies was the length of the task, as in Study 5 participants interacted with 1-2 more persons, before getting to the response with the target. Consequently, the norms were likely not as salient anymore when finally responding to the target, resulting in weaker overall effects. This effect of the saliency of the norms was the focus of a study by Matias (2019), where the community norms were explicitly displayed with a “sticky comment” atop the forum to newcomers in a large-scale online science-discussion community. The study found that simply showing the rules to new participants prominently significantly increased both rule compliance and participation rates, demonstrating the powerful impact of making norms salient in digital environments. This is not just a methodological

point. Rather, all insights regarding the impact of codes of conduct, whether positive or negative, are dependent on this lever of saliency. Generally, CoCs also become more binding if they are enforced more reliably, a problem that poses plenty of challenges at scale (e.g., Riberio et al., 2022).

8.3. Methodological Implications

Within this contribution, we introduced a novel paradigm, ArtEval, which has many advantages over previous paradigms investigating norm-enforcement. To our knowledge, this is the first paradigm that allows us to investigate both ostracism and confrontation at the same time. Especially in terms of ecological validity, this variety is a crucial step up, since, in the real world, whether online or offline, it is rarely the case that one can *only* decide to ostracize or not or *only* decide to confront or not. Furthermore, considering the increasing capabilities of scalable nuanced text data-analyses with Large-Language-Models (LLMs, e.g. Rathje et al., 2023), the written component of the norm-enforcement can allow for even more fine-grained analyses. The generally even attractiveness of both ostracism and confrontation in response to the rude behavior (see *Studies 1-2*), allows to introduce experimental factors that can tip that balance in favor of either response. Thus far, there is little research on what factors determine how people respond to norm-violations (e.g., Molho et al., 2017, 2020), and even less experimental research. ArtEval could be used as a starting point for further inquiries in this area of research (for access to the research materials, see the OSF repository: https://osf.io/nk5a8/?view_only=fc99ab6692e5431dbdf9069692cab6cd). While the paradigm can generally also be easily adapted when it comes to other characteristics like the specific topic discussed in the forum, or the specific shape that the norm-violation takes on, this can alter the beneficial characteristics of the baseline-setup (see *Study 4*).

8.4. Limitations

The interpretation of our operationalization of ostracism is non-trivial, as the motives for the observed behavior could have been manifold. For one, the first study showed that even a potentially unpleasant conversation partner can spark modest rates of ostracism, despite not violating any norms.

Consequently it is likely that not all of the ostracism in our studies is an effort to subtly punish the target. Also, as was mentioned in the discussion of Study 1, there could also simply be interpersonal differences of what deserves to be punished (Carlsmith et al., 2002). But again, the shifts in the behavioral pattern that causally follow from the manipulation of the CoC allow for confidence that this was the case for a proper share of the participants. Lastly, choosing to “skip and exclude” could also just result from lack of motivation and wanting to finish the study as quickly as possible. While this plausibly holds true for a share

9. Conclusion

The impact of rules of conduct on norm-enforcement behavior in online forums is an important question, to understand and guide the increasing amount of social context in virtual social spaces. Here, we present evidence that rules of conduct that discourage conflict increase the likelihood of ostracism as a means of norm-enforcement, while norms that encourage conflict increase the likelihood of confrontation as a means of norm-enforcement.

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.

References

- Abbink, K., Gangadharan, L., Handfield, T., & Thrasher, J. (2017). Peer punishment promotes enforcement of bad social norms. *Nature Communications*, 8(1), 609. <https://doi.org/10.1038/s41467-017-00731-0>
- Archer, J., & Coyne, S. M. (2005). An Integrated Review of Indirect, Relational, and Social Aggression. *Personality and Social Psychology Review*, 9(3), 212–230. https://doi.org/10.1207/s15327957pspr0903_2

- Ashburn-Nardo, L., Blanchar, J. C., Petersson, J., Morris, K. A., & Goodwin, S. A. (2014). Do you say something when it's your boss? The role of perpetrator power in prejudice confrontation. *Journal of Social Issues*, 70(4), 615–636. <https://doi.org/10.1111/josi.12082>
- Auxier, B., & Anderson, M. (2021). Social media use in 2021. *Pew Research Center*, 1, 1–4. <https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/>
- Axelrod, R. (1986). An evolutionary approach to norms. *The American Political Science Review*, 80(4), 1095–1111. <https://doi.org/10.2307/1960858>
- Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, 7(1), 1–6. <https://doi.org/10.1038/ncomms13327>
- Balliet, D., & Van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, 139(5), 1090–1112. <https://doi.org/10.1037/a0030939>
- 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>
- Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, 149(4), 746–756. <https://doi.org/10.1037/xge0000673>
- 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>
- Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2), 284–299. <https://doi.org/10.1037/0022-3514.83.2.284>
- Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction*, 32, 1–25. <https://doi.org/10.1145/3274301>
- Cialdini, R., Demaine, L., Sagarin, B., Barrett, D., Rhoads, K., & Winter, P. (2006). Managing social norms for persuasive impact. *Social Influence*, 1(1), 3–15. <https://doi.org/10.1080/15534510500181459>

- Cialdini, R.B., Kallgren, C.A., & Reno, R.R. (1991). A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. *Advances in Experimental Social Psychology*, 24, 201-234. [https://doi.org/10.1016/S0065-2601\(08\)60330-5](https://doi.org/10.1016/S0065-2601(08)60330-5)
- Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. *The handbook of social psychology*, 4, 151–192.
- Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, 14(2), 81-100. <https://doi.org/10.1037/a0015914>
- 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>
- Dickter, C. L., & Newton, V. A. (2013). To confront or not to confront: Non-targets' evaluations of and responses to racist comments. *Journal of Applied Social Psychology*, 43, E262–E275. <https://doi.org/10.1111/jasp.12022>
- Drury, B. J., & Kaiser, C. R. (2014). Allies against Sexism: The Role of Men in Confronting Sexism. *Journal of Social Issues*, 70(4), 637–652. <https://doi.org/10.1111/josi.12083>
- Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, 302(5643), 290–292. <https://doi.org/10.1126/science.1089134>
- Eriksson, K., Andersson, P. A., & Strimling, P. (2017). When is it appropriate to reprimand a norm violation? The roles of anger, behavioral consequences, violation severity, and social distance. *Judgment and decision making*, 12(4), 396–407. <https://doi.org/10.1017/S1930297500006264>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Feagin, J. R. (1991). The continuing significance of race: Antiracist discrimination in public places. *American Sociological Review*, 56(1), 101–116. <https://doi.org/10.2307/2095676>
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785–791. <https://doi.org/10.1038/nature02043>
- 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>
- Freedman, G., Hales, A. H., Powell, D. N., Le, B., & Williams, K. D. (2022). The role of gender and safety concerns in romantic rejection decisions. *Journal of Experimental Social Psychology*, 102, 104368. <https://doi.org/10.1016/j.jesp.2022.104368>

- Gibson, S. (2019). Obedience without orders: Expanding social psychology's conception of 'obedience'. *British Journal of Social Psychology*, 58(1), 241–259. <https://doi.org/10.1111/bjso.12272>
- Grace, T., Larson, I., & Salen, K. (2022). Policies of Misconduct: A Content Analysis of Codes of Conduct for Online Multiplayer Games. *Proceedings of the ACM on Human-Computer Interaction*, 6, 1 - 23. <https://doi.org/10.1145/3549513>
- Grey, T. (1991). Civil Rights Vs. Civil Liberties: The Case of Discriminatory Verbal Harassment. *Social Philosophy and Policy*, 8, 81 - 107. <https://doi.org/10.1017/S026505250000114X>
- Hales, A. H., Ren, D., & Williams, K. D. (2017). Protect, correct, and eject: Ostracism as a social influence tool. *The Oxford Handbook of Social Influence*, 205–217. <https://doi.org/10.1093/oxfordhb/9780199859870.013.26>
- Hayes, R. A., Wesselmann, E. D., & Carr, C. T. (2018). When Nobody “Likes” You: Perceived Ostracism Through Paralinguistic Digital Affordances Within Social Media. *Social Media + Society*, 4(3). <https://doi.org/10.1177/2056305118800309>
- Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of theoretical biology*, 208(1), 79-89. <https://doi.org/10.1006/jtbi.2000.2202>
- 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>
- Herrera Enríquez, M. D. C., Herrera, A., & Expósito Jiménez, F. (2018). To confront versus not to confront: Women's perception of sexual harassment. *European Journal of Psychology Applied to Legal Context*, 10(1), 1–7. <https://doi.org/10.1016/j.ejpal.2017.04.002>
- Hunter, R., De La Haye, K., Murray, J., Badham, J., Valente, T., Clarke, M., & Kee, F. (2019). Social network interventions for health behaviours and outcomes: A systematic review and meta-analysis. *PLoS Medicine*, 16(9), e1002890. <https://doi.org/10.1371/journal.pmed.1002890>
- Johnson, D. W., & Johnson, R. T. (2009). Energizing learning: The instructional power of conflict. *Educational Researcher*, 38(1), 37–51. <https://doi.org/10.3102/0013189X08330540>
- Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, 530(7591), 473–476. <https://doi.org/10.1038/nature16981>
- Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, 118(1), 57–88. <https://doi.org/10.1037/pspi0000186>
- Kaiser, C. R., & Miller, C. T. (2001). Stop complaining! The social costs of making attributions to discrimination. *Personality and Social Psychology Bulletin*, 27(2), 254–263. <https://doi.org/10.1177/0146167201272010>

- Kameda, T., Takezawa, M., & Hastie, R. (2003). The Logic of Social Sharing: An Evolutionary Game Analysis of Adaptive Norm Development. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, 7, 2–19. https://doi.org/10.1207/S15327957PSPR0701_1
- Kang, K. (2017). Anonymity and Interaction in an Online Breast Cancer Social Support Group. *Communication Studies*, 68(4), 403 - 421. <https://doi.org/10.1080/10510974.2017.1340902>
- Kenntemich, C., Brückner-Collet, D. O. I., & Rudert, S. C. (2024a). *The Global Village's Local Rules: Cross-Cultural Differences in Online Norm-Enforcement*. Manuscript submitted for publication.
- Kenntemich, C., Büttner, C.M., & Rudert, S.C. (2024b). The Pursuit of Approval: Social Media Users' Decreased Posting Latency Following Online Exclusion as a Form of Acknowledgement-Seeking Behavior. *Personality & Social Psychology Bulletin*. <https://doi.org/10.1177/01461672241297824>
- 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>
- Kerr, N. L., Rumble, A. C., Park, E. S., Ouwerkerk, J. W., Parks, C. D., Gallucci, M., & van Lange, P. A. M. (2009). “How many bad apples does it take to spoil the whole barrel?”: Social exclusion and toleration for bad apples. *Journal of Experimental Social Psychology*, 45(4), 603–613. <https://doi.org/10.1016/j.jesp.2009.02.017>
- Kurzban, R., & Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological bulletin*, 127(2), 187–208. <https://doi.org/10.1037/0033-2909.127.2.187>
- Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2), 434–443. <https://doi.org/10.1016/j.chb.2011.10.014>
- Lu, D., Li, Q., & Liao, S. (2012). A graph-based action network framework to identify prestigious members through member's prestige evolution. *Decis. Support Syst.*, 53(1), 44-54. <https://doi.org/10.1016/j.dss.2011.12.003>
- Maier-Rigaud, F. P., Martinsson, P., & Staffiero, G. (2010). Ostracism and the provision of a public good: Experimental evidence. *Journal of Economic Behavior & Organization*, 73(3), 387–395. <https://doi.org/10.1016/j.jebo.2009.11.001>
- Mallett, R. K., & Wagner, D. E. (2011). The unexpectedly positive consequences of confronting sexism. *Journal of Experimental Social Psychology*, 47(1), 215–220. <https://doi.org/10.1016/j.jesp.2010.10.001>

- Martin, J. W., Jordan, J. J., Rand, D. G., & Cushman, F. (2019). When do we punish people who don't? *Cognition*, 193, 104040. <https://doi.org/10.1016/j.cognition.2019.104040>
- Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20), 9785–9789. <https://doi.org/10.1073/pnas.1813486116>
- Molho, C., Tybur, J. M., Güler, E., Balliet, D., & Hofmann, W. (2017). Disgust and anger relate to different aggressive responses to moral violations. *Psychological Science*, 28(5), 609–619. <https://doi.org/10.1177/0956797617692000>
- Molho, C., Tybur, J. M., Van Lange, P. A. M., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications*, 11(1), 1–9. <https://doi.org/10.1038/s41467-020-17286-2>
- Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of personality and social psychology*, 83(5), 1029–1050. <https://doi.org/10.1037/0022-3514.83.5.1029>
- Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, 39, 629–649. <https://doi.org/10.1007/S11109-016-9373-5>
- Muthén, L. K., & Muthén, B. O. (2002). How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8
- Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves?. *Journal of Public Economics*, 92(1-2), 91–112. <https://doi.org/10.1016/j.jpubeco.2007.04.008>
- Nordgren, L. F., Banas, K., & MacDonald, G. (2011). Empathy gaps for social pain: Why people underestimate the pain of social suffering. *Journal of personality and social psychology*, 100(1), 120–128. <https://doi.org/10.1037/a0020938>
- O'Reilly, J., Robinson, S. L., Berdahl, J. L., & Banki, S. (2015). Is negative attention better than no attention? The comparative effects of ostracism and harassment at work. *Organization Science*, 26(3), 774–793. <https://doi.org/10.1287/orsc.2014.0900>
- Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113(3), 566–571. <https://doi.org/10.1073/pnas.1514483113>
- Paskuda, M., & Lewkowicz, M. (2017). Anonymity interacting with participation on a Q&A site. *AI & SOCIETY*, 32, 369–381. <https://doi.org/10.1007/s00146-016-0660-9>

- Rashidi, Y., Kapadia, A., Nippert-Eng, C., & Su, N. M. (2020). "It's easier than causing confrontation": Sanctioning Strategies to Maintain Social Norms and Privacy on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1–25. <https://doi.org/10.1145/3392827>
- Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjeh, R., Robertson, C., & Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 121(34), e2308950121. <https://doi.org/10.31234/osf.io/sekf5>
- Redhead, D., Dhaliwal, N., & Cheng, J. T. (2021). Taking charge and stepping in: Individuals who punish are rewarded with prestige and dominance. *Social and Personality Psychology Compass*, 15(2), e12581. <https://doi.org/10.1111/spc3.12581>
- Ribeiro, M., Cheng, J., & West, R. (2022). Automated Content Moderation Increases Adherence to Community Guidelines. *Proceedings of the ACM Web Conference 2023*. <https://doi.org/10.1145/3543507.3583275>
- Robinson, S. L., O'Reilly, J., & Wang, W. (2013). Invisible at Work: An Integrated Model of Workplace Ostracism. *Journal of Management*, 39(1), 203–231. <https://doi.org/10.1177/0149206312466141>
- Rudert, S. C., Hales, A. H., & Buttner, C. M. (2021). Stay out of our office (vs. our pub): Target personality and situational context affect ostracism intentions*. *Journal of Experimental Social Psychology*, 95, 104142. <https://doi.org/10.1016/j.jesp.2021.104142>
- Rudert, S. C., Hales, A. H., Greifeneder, R., & Williams, K. D. (2017). When Silence Is Not Golden: Why Acknowledgment Matters Even When Being Excluded. *Personality and Social Psychology Bulletin*, 43(5), 678–692. <https://doi.org/10.1177/0146167217695554>
- 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>
- Strimling, P., & Eriksson, K. (2014). Regulating the regulation: Norms about punishment. In P. A. M. Van Lange, B. Rockenbach, & T. Yamagishi (Eds.), *Reward and punishment in social dilemmas* (pp. 52–69). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199300730.003.0004>
- Swim, J. K., & Hyers, L. L. (1999). Excuse me—What did you just say?!: Women's public and private responses to sexist remarks. *Journal of Experimental Social Psychology*, 35(1), 68–88. <https://doi.org/10.1006/jesp.1998.1370>
- Villatoro, D., Andrighetto, G., Brandts, J., Nardin, L. G., Sabater-Mir, J., & Conte, R. (2014). The norm-signaling effects of group punishment: combining agent-based simulation and laboratory experiments. *Social science computer review*, 32(3), 334–353. <https://doi.org/10.1177/0894439313511396>

- Wellek, S. (2002). *Testing statistical hypotheses of equivalence*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781420035964>
- Williams, K. D. (2007). Ostracism: The kiss of social death. *Social and Personality Psychology Compass*, 1(1), 236–247. <https://doi.org/10.1111/j.1751-9004.2007.00004.x>
- Williams, K. D. (2009). Ostracism: A temporal need-threat model. *Advances in Experimental Social Psychology*, 41, 275–314. [https://doi.org/10.1016/S0065-2601\(08\)00406-1](https://doi.org/10.1016/S0065-2601(08)00406-1)
- Williams, K. D., & Nida, S. A. (2022). Ostracism and social exclusion: Implications for separation, social isolation, and loss. *Current Opinion in Psychology*, 47, 101353. <https://doi.org/10.1016/j.copsyc.2022.101353>

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
1–19

© The Author(s) 2024



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01461672241297824
journals.sagepub.com/home/pspb



Christoph Kenntemich^{1*}, Christiane M. Büttner^{*,2},
and Selma C. Rudert¹

Abstract

How do individuals behave after the sting of social exclusion on social media? Previous theorizing predicts that, after experiencing exclusion, individuals either engage in activities that reconnect them with others, or, they withdraw from the context. We analyzed data from Twitter ($k = 47,399$ posts; $N = 2,000$ users) and Reddit ($k = 58,442$ posts; $N = 2,000$ users), using relative (un)popularity of users' own posts (i.e., receiving fewer Likes/upvotes than usual) as an indicator of social exclusion. Both studies found no general increase or decrease in posting latency following exclusion. However, the latency of behaviors aimed at connecting with *many others* decreased (i.e., posting again quickly), and the latency of behaviors aimed at connecting with *specific others* increased (i.e., commenting or mentioning others less quickly). Our findings speak in favor of acknowledgment-seeking behavior as a reaction to social exclusion that may be specific to social media contexts.

Keywords

social exclusion, ostracism, social media, behavioral coping

Received October 13, 2023; revised accepted October 19, 2024

Social media has become an integral part of human life, including the way that social interactions are lived (e.g., Büttner et al., 2023; Reich et al., 2018). Social media offers novel ways of connecting with other people, simplifying social interactions into paralinguistic digital affordances such as Likes, hearts, “+1,” and emojis (e.g., Hayes et al., 2016). However, these novel forms of social interaction also pave the way for negative experiences, such as novel forms of exclusion (i.e., feeling ignored and excluded by others on social media, e.g., Williams, 2009; Wolf et al., 2015). For instance, not getting a satisfactory number of paralinguistic digital affordances such as Likes in response to a social media post is perceived as social exclusion (e.g., Lutz & Schneider, 2020; Reich et al., 2018), even in fictional social media networks (e.g., Wolf et al., 2015). Following these observations, we use the relative (un)popularity of users' own posts (i.e., receiving fewer Likes/upvotes than usual) as an indicator of social exclusion. Social exclusion has highly aversive consequences for its targets both in offline and online contexts, such as threatening needs of belonging, control, self-esteem, and meaningful existence (e.g., Büttner & Rudert, 2022; Williams, 2009; Wolf et al., 2015). If not remedied, social exclusion fosters depression and suicidality in the long run (e.g., Büttner & Greifeneder, 2024; Chen et al., 2020). Importantly, social

exclusion is distinct from experiencing negative attention, such as rejection or bullying: Social exclusion is the absence of attention and is more emotionally aversive than rejection (Rudert et al., 2017) or bullying (Zadro et al., 2005). Social exclusion represents being rendered meaningless, a severely threatening experience compared with receiving negative attention where at least the target receives some attention (e.g., Rudert et al., 2017; Williams & Nida, 2009). On social media, not receiving a usual amount of Likes, an exclusionary experience, is equally threatening to needs for meaningful existence and control as receiving Dislikes, that is, negative attention (but receiving Dislikes threatened needs for belonging and self-esteem more strongly than social exclusion, Lutz & Schneider, 2020).

To cope with social exclusion and recover their threatened needs, individuals can engage in various behaviors, such as

¹University of Kaiserslautern-Landau, Germany

²University of Basel, Switzerland

*C.K. and C.M.B. shared first-author authorship.

Corresponding Author:

Christoph Kenntemich, Social Environmental and Economic Psychology,
University of Kaiserslautern-Landau, Fortstr. 7, Landau D-4055, Germany.
Email: c.kenntemich@rptu.de

seeking reconnection with others, aggressing against them, or withdrawing from social interactions altogether (e.g., Maner et al., 2007; Ren et al., 2018, 2020). Social media also offers features that may help excluded individuals cope with their experience, such as posting content or leaving Likes, Dislikes, or comments under other users' posts (e.g., Karahanna et al., 2018; Lutz et al., 2022, 2023). A shared problem of previous research into behavioral consequences of social exclusion, in the real world and in social media contexts, is that the often-used laboratory studies come at a cost of decreased external validity of participants' assessed behaviors (see, for example, Baumeister et al., 2007). We propose to close this gap with observational data from two social media platforms: Twitter (at the time of data mining, now "X") and Reddit. Twitter is a micro-blogging social media platform where users post *tweets* (short texts) and interact with other users through *favorites* (liking another user's post), *retweets* (reposting another user's post), *mentions* (addressing another user by linking their profile name in direct response or in a comment), and *comments* that respond to another user's post (e.g., Karahanna et al., 2018). Reddit is a social media platform focused on content sharing and discussion among users. Reddit users may post links, videos, news, pictures, or texts in *subreddits* (forums ordered by topic) where other users can respond via *comments* and interact with *upvotes* or *downvotes* to the post, indicating popularity of the content (e.g., Proferes et al., 2021).

We formulate three hypotheses on how individuals may behave after being socially excluded on social media: Users may behave to reconnect with *many* others and behave to reconnect with *specific individuals*, or they may withdraw from the situation. We review arguments and empirical evidence for each hypothesis in the following.

Hypothesis 1: The "Reconnection with Many" Hypothesis.

Social exclusion disconnects individuals from others, causing them to seek reconnection (e.g., Eck et al., 2016; Williams, 2009). Social media is a unique coping tool in that it offers the possibility to connect with *many* individuals simultaneously, for instance by posting content for everyone to see (e.g., Karahanna et al., 2018). By posting content, social media users may express themselves and exert control over how they want others to perceive them (e.g., as highly knowledgeable, popular, or connected, see, for example, Büttner et al., 2023). Posting content may also be an effective tool in restoring relatedness and other needs threatened by social exclusion (e.g., Lutz et al., 2022). We conceptualize posting to reconnect with *many* others as behaving to be *acknowledged* rather than to be *liked* after exclusion on social media. Relating to Williams's (2009) Temporal Need Threat Model, exclusion on social media, a context where individuals seek acknowledgment from others (Karahanna et al., 2018), threatens power and provocation needs (i.e., needs for control and for being acknowledged in one's existence by others, Williams,

2009) more than inclusionary needs (i.e., needs for belonging and self-esteem, Williams, 2009), in line with previous experimental findings (Lutz & Schneider, 2020). Thus, with a mass social media audience, the quality of exclusion encapsulated by few Likes could lead to a specific deficit in acknowledgment that is qualitatively different from other exclusions. To re-afford these needs, actions that are aimed at being acknowledged by many others might be more effective for individuals' coping than actions aimed at reconnecting with specific others. In summary, users who just experienced exclusion on social media may post content faster to a large audience to increase their chances of reconnecting with many others.

Hypothesis 2: The "Reconnection with Specific Individuals" Hypothesis. Instead of reconnecting with a vast number of individuals, the Reconnection with Specific Individuals Hypothesis holds that excluded individuals may turn to specific persons for reconnection. Specifically addressing another user in a comment, or mentioning them, might be more effective to prompt an answer from that person and thus connect with them, than posting content and hoping for one out of an unspecified throng of people to respond. On Twitter, a mention is a post that is either made in direct response to another user's post (i.e., @user followed by a comment under that user's post) or that otherwise addresses another user directly (i.e., @user in a post). A retweet on Twitter echoes another user's post to the reposting user's followers. Both mentions and retweets may evoke reciprocal communication: Reciprocity is a strong interaction norm that holds that individuals employ a "mutually contingent exchange of gratifications" (Gouldner, 1960, p. 161). Consistent with this notion, Twitter users retweet others' content out of *reciprocity motivation*, that is, because they expect a message in return, or because their content had been previously retweeted (M. Lee, Kim, & Kim, 2015). Thus, the likelihood of receiving an answer from the user that one has mentioned or retweeted may be higher because of reciprocity norms than the likelihood that others respond to a post that is not tailored to one specific individual. Importantly, even being included by one single person effectively alleviates the negative impact of social exclusion (e.g., DeWall et al., 2010). Perceived chances of re-inclusion also shape excluded individuals' behavioral reactions (Cuadrado et al., 2015). Chances of re-inclusion may be higher when approaching specific others versus a large group of social media users, which may cause excluded individuals to choose the former strategy over the latter. Moreover, commenting on other users' posts or mentioning other users and their content may already boost one's relatedness needs (i.e., meta-voicing, Karahanna et al., 2018). In conclusion, users who experienced exclusion may engage in reconnection behavior aimed at *specific individuals* such as commenting or sharing other users' content.

Hypothesis 3: The Withdrawal Hypothesis.

Experimental research documents large effects of social exclusion on withdrawal from others (e.g., Ren et al., 2020). These effects may seem contradictory given the evidence reviewed above on reconnection efforts after exclusion. However, by withdrawing, socially excluded individuals may protect themselves against future negative experiences in the same environment (e.g., Ren et al., 2020). Social media is a context that is especially easy to leave without immediate negative consequences (as opposed to, leaving a work meeting, or a school lesson). Therefore, it could also be that, following social media exclusion, individuals withdraw and refrain from using social media for some time to seek connections elsewhere.

The Present Research

Understanding how users respond to unfavorable reception in social media environments is a pressing research question. Using data from Twitter and Reddit, we test whether the users' time until the next post or comment increased or decreased following exclusion, in line with the idea that posting and commenting represent immediate coping responses to exclusion.

Operationalizations

We analyzed users' online behaviors following social exclusion, defined by a lack of popularity of users' posts compared with the usual popularity of a user's post (see Lutz & Schneider, 2020; Reich et al., 2018). Particularly, we focused on the latency until users' next post or comment after an unpopular post. We operationalized shorter latency (less time) until a user's next post as indicative of attempts to *reconnect with many* other social media users, shorter latency until a user's next mention, retweet (Study 1, Twitter), or comment (Study 2, Reddit) as indicative of attempts to *reconnect with specific individuals*, and, finally, we operationalized longer latency (more time) until a user's next post as indicative of *withdrawal* from social media.

Unpopular Versus Regularly Popular Posts

Importantly, we hypothesize that the effects are primarily influenced by social exclusion. If social exclusion is the driving factor, the behavioral differences should be more pronounced when contrasting an unusually unpopular post with a regularly popular one, as opposed to comparing an exceptionally popular post with one of average popularity. That is, because we expect the effects to be driven by social exclusion, all of the predicted relations are not necessarily linear; we only predict differences between a particularly low number of paralinguistic digital affordances (i.e., in the lower third of the distribution, what we define as a point of minimal

social disappointment) versus a not-low number of paralinguistic digital affordances (i.e., the two upper thirds of the distribution). We contrast the lowest third with the two upper thirds because unexpectedly popular posts (i.e., the upper third) are not the focus of this contribution. The choice of the lower third of the distribution as the point of minimal social disappointment was data-driven since the tertile split fit the data better than a median split, quartile split, or a cubic predictor (see Open Science Framework [OSF] Supplement 1a—l. 419ff). However, a continuous quadratic predictor was also a good fit compared with a baseline model (see OSF Supplementary model fits). Therefore, we explore quadratic relationships in exploratory fashion in both studies.

Method Statement

Open Science Practices. We report all data exclusions, data manipulations, and measures. All code for data mining, data preparation, and statistical analyses, as well as the data for both studies is available via the OSF (https://osf.io/h62gz/?view_only=13360594eb654ab4a690ebdd11a4da00). We preregistered all hypotheses, sample sizes, exclusion criteria, and analysis plans on AsPredicted (Study 1: <https://aspredicted.org/kt8v-jrhx.pdf>; Study 2: <https://aspredicted.org/ndvn-276b.pdf>).

Data Mining. As an indicator of social exclusion, we focused on the *relative* popularity of users' *own content*. Many experimental exclusion paradigms rely on social comparison with other individuals (e.g., receiving fewer Likes than other social media users, Wolf et al., 2015). Here, the users' own previous posting experiences serve as the reference point, comparing users' post popularity to their own previous posts. To create our predictor variable, we excluded all posts in which users mentioned others, or retweeted other people's content on Twitter (Study 1), and we only analyzed posts as a predictor (as opposed to comments on others' posts) on Reddit (Study 2). To allow for intraindividual comparisons, we excluded individuals with less than 10 posts in total. To ensure the normality of our data and to focus on the range of theoretical interest, as preregistered, we excluded uncommonly popular posts from our sample. To do so, first, we transformed the independent variable (i.e., in Study 1: $X = 2 \times \sqrt{\text{favorite count}}$, in Study 2: $X = 2 \times \sqrt{\text{score}}$). Next, following a median absolute derivation approach (e.g., Howell, 2005), we defined uncommonly popular posts as exceeding a set robust threshold for count data, that is, $X > \text{Median}(X) + 10$. While a more typical threshold for extreme value identification is set at 3 units above the median, that is, $X > \text{Median}(X) + 3$, we adopted a more liberal criterion to account for potential low gamma of the distribution and exclude extremely popular posts exclusively. This excluded roughly the most popular 13% of posts in the Twitter sample and the most popular 30% in the Reddit sample.¹ This exclusion criterion was not preregistered for Study 1 but was

preregistered for Study 2. In exploratory fashion, we explore the analyses again in the full data set, without excluding extremely popular posts (Tables 1 and 4).

Analytical Strategy

Mixed Effect Modeling. We conducted mixed effect modeling using the lme4 (Bates et al., 2015) and glmmTMB (Brooks

et al., 2017) R-packages. We aimed for maximal linear mixed effect models (LMEMs), including random intercepts and random slopes, which are considered the gold standard (Barr et al., 2013). For exceptions when the maximal LMEM did not converge, we report the results from random-intercepts-only models. All reported *p* values are two-tailed.

The formula of the multilevel models is as follows:

$$\text{Level 1 (Within Individuals): Latency}_{ij} = \beta_{0j} + \beta_{1j}(\text{Fav}_{ij}) + \beta_{2j}(\text{Fav}_{ij} \times \text{Tertile}_{ij}) + r_{ij}$$

Latency_{ij} : Log-transformed latency to the next own post for post *i* of user *j*.

Fav_{ij} : Z-score of the Anscombe-transformed favorites for post *i* of user *j*.

Tertile_{ij} : Whether post *i* is in the lowest tertile of Favorites for user *j*.

β_{0j} : Intercept for user *j*.

β_{1j} : Slope of the effect of favorites on latency for user *j*.

β_{2j} : Slope for interaction between favorites and being in the lowest tertile for user *j*.

r_{ij} : Residual for post *i* of user *j*.

$$\text{Level 2 (Between Individuals): } \beta_{0j} = \gamma_{00} + u_{0j};$$

$$\beta_{1j} = \gamma_{10} + u_{1j};$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

γ_{00} : Overall intercept (latency when favorites and lowest tertile are both 0).

γ_{10} : Slope of favorites' effect on latency for user *j*.

γ_{20} : Effect of being in the lowest tertile of favorites on the slope of favorites.

u_{0j} , u_{1j} , and u_{2j} : Random effects capturing the deviation of user *j* from the average intercept (γ_{00}), slope for Favorites (γ_{10}), and the slope for the interaction term (γ_{20}).

Variable Transformations. Latencies were log-transformed, that is, log (latency + 1) to improve linearity. Twitter favorites and Reddit scores were Anscombe-transformed to increase the normality of the highly skewed distribution (Anscombe, 1948) and standardized within individual (i.e., centering within cluster) to produce individual-level estimates (Enders & Tofighi, 2007). Degrees of freedom, test statistics, and *p* values were derived from Satterthwaite approximations in the lmerTest package (Kuznetsova et al., 2017). The exact formulas for all variable transformations can be found at the end of each of the R-Scripts (OSF).

Interpretation of Effect Sizes. We provide log-transformed regression coefficients and 95% confidence intervals (95% CI). Since all dependent variables were log-scaled, this coefficient can be interpreted as follows: For a post that is one standard deviation less popular, the decrease in time until the next post translates to a factor of e^{-b} . So for $b = 0.1$: $e^{-0.1} \approx 0.905$, corresponding to a 9.52% decrease.

We report percentage decreases as an interpretation of the effect (Tables 1 and 4).

Study 1

Study 1 explored responses to exclusion online with competing hypotheses on behavioral responses. We test the following hypotheses in Study 1:

Hypothesis 1: "Reconnection with Many": The relative popularity of a tweet, operationalized as the number of favorites of a tweet relative to the usual number of favorites that a user receives, is positively associated with the time until posting the next tweet (posting latency).

Hypothesis 2: "Reconnection with Specific Individuals": The relative popularity of a tweet is positively associated with (a) the time until the next mention (mentioning latency) and (b) the time until the next retweet (retweeting latency).

We preregistered that these relations are not necessarily linear, as we predict differences between a low number of favorites (exclusion) and a not-low count of favorites, and, therefore, no differences between an average amount of favorites and an above-average amount of favorites. In Study 1, we operationalized low number of favorites versus not-low number of favorites by exploratorily dividing the distribution into the lowest third versus the two upper thirds.

As a competing hypothesis to Hypotheses 1 and 2, we preregistered:

Hypothesis 3: "Withdrawal Hypothesis": The relative popularity of a tweet is (a) negatively associated with time until posting the next tweet (posting latency), (b) time until the next mention (mentioning latency), and (c) time until the next retweet (retweeting latency).

Method

Social Media Data. We sampled the latest 150 posts for a random sample of users and filtered them to maximize number of relevant data points per user and share of individual persons (as opposed to, organizations or companies; we excluded data from verified users² and users that did not have a name

Table 1. Results of the LMEMs in Study 1, With Filtered (i.e., Excluding Extremely Popular Posts) and Unfiltered (i.e., Full Data Set) Data.

Data	Latency variable	Main effect b [95% CI]	Main effect—% $SD(-1\ SD)$	Lower third b [95% CI]	Lower third—% $(+/-1\ SD)$	Upper two thirds b [95% CI]	Upper two thirds—% $(+/-1\ SD)$	Quadratic term b [95% CI]	Fixed effects (SE)	Random effects (SE)
Filtered	Next post	0.09 [0.07, 0.11], $p < .001$	+09.64%/-8.79%	0.16 [0.12, 0.20], $p < .001$	+17.22%/-14.69%	0.05 [0.02, 0.07], $p = .001$	+4.77%/-4.56%	-0.03 [-0.04, -0.01], $p < .001$	$\beta_0 = 11.93$ (0.04), $p < .001$; $\beta_1 = 0.05$ (0.01), $p = .001$; $\beta_2 = 0.11$ (0.03), $p < .001$	$u_g = 2.71$; $u_{ij} = 0.03$; $Cov(u_g, u_{ij}) = -0.26$; $r_g = 3.47$
	Next mention	-0.15 [-0.17, -0.12], $p < .001$ [†]	-13.50%/-15.60%	-0.06 [-0.11, 0.02], $p = .054$	-6.17%/-6.58%	-0.20 [-0.23, -0.16], $p < .001$	-18.03%/-22.00%	-0.03 [-0.05, -0.02], $p < .001$ [†]	$\beta_0 = 11.03$ (0.05), $p < .001$; $\beta_1 = -0.20$ (0.02), $p < .001$; $\beta_2 = 0.14$ (0.03), $p < .001$	$u_g = 3.60$; $u_{ij} = 0.06$; $Cov(u_g, u_{ij}) = -0.01$; $r_g = 5.11$
	Next retweet	0.04 [0.02, 0.06], $p < .001$	+3.67%/-3.54%	0.08 [0.04, 0.12], $p < .001$	+8.44%/-7.78%	0.01 [-0.02, 0.04], $p = .666$	+0.63%/-0.63%	-0.01 [-0.03, -0.00], $p = .049$	$\beta_0 = 11.94$ (0.04), $p < .001$; $\beta_1 = 0.06$ (0.01), $p = .666$; $\beta_2 = 0.07$ (0.03), $p = .005$	$u_g = 2.98$; $u_{ij} = 0.02$; $Cov(u_g, u_{ij}) = -0.15$; $r_g = 3.43$
Unfiltered	Next post	0.08 [0.06, 0.10], $p < .001$	+8.55%/-7.87%	0.16 [0.11, 0.21], $p < .001$	+17.20%/-14.67%	0.05 [0.03, 0.08], $p < .001$	+5.49%/-5.21%	-0.03 [-0.04, -0.02], $p < .001$ [†]	$\beta_0 = 11.80$ (0.04), $p < .001$; $\beta_1 = 0.05$ (0.01), $p < .001$; $\beta_2 = 0.11$ (0.03), $p < .001$	$u_g = 1.56$; $u_{ij} = 0.002$; $Cov(u_g, u_{ij}) = -0.96$; $r_g = 4.63$
	Next mention	-0.20 [-0.22, -0.18], $p < .001$ [†]	-18.13%/-22.14%	-0.18 [-0.23, -0.12], $p < .001$ [†]	-16.16%/-19.27%	-0.21 [-0.24, -0.18], $p < .001$	-18.67%/-22.95%	0.00 [-0.01, 0.02], $p = .484$	$\beta_0 = 10.95$ (0.05), $p < .001$; $\beta_1 = -0.21$ (0.02), $p < .001$; $\beta_2 = 0.03$ (0.04), $p = .848$	$u_g = 1.82$; $r_g = 6.86$
	Next retweet	0.02 [-0.002, 0.04], $p = .072$	+1.92%/-1.88%	0.12 [0.07, 0.16], $p < .001$	+12.39%/-11.03%	-0.02 [-0.05, 0.005], $p = .108$	-2.10%/-2.15%	-0.02 [-0.03, -0.01], $p < .001$ [†]	$\beta_0 = 11.98$ (0.04), $p < .001$; $\beta_1 = -0.02$ (0.01), $p = .108$; $\beta_2 = 0.14$ (0.03), $p < .001$	$u_g = 1.74$; $r_g = 4.71$

Note. Significant effects are bold-faced. [†] Random-intercept only model. Parameters: intercept (β_0), predictor favorite count (β_1), discretizing factor (β_2), intercept variance (u_g), slope variance (u_{ij}), covariance (u_g, u_{ij}), residual variance (η_j).

Table 2. Within-Study Replication Analyses for Studies 1 and 2.

Study	Variable	Share significant— main effect	Effect range — main effect	Share significant— lowest tertile	Effect range— lowest tertile	Share significant— upper two thirds	Effect range— upper two thirds	Share significant— quadratic	Effect range— quadratic
1	Latency next post	5/5 (100%)	0.07 to 0.10	5/5 (100%)	0.12 to 0.18	3/5 (60%)	0.02 to 0.05	4/5 (80%)	-0.04 to -0.02
	Latency next retweet ^a	5/5 (100%)	0.03 to 0.04	4/5 (80%)	0.04 to 0.09	0/5 (0%)	0.00 to 0.03	1/5 (80%)	-0.02 to -0.01
	Latency next mention	5/5 (100%)	-0.16 to -0.11	2/5 (40%)	-0.08 to -0.03	5/5 (100%)	-0.25 to -0.14	4/5 (80%)	-0.05 to -0.01
2	Latency next post (any subreddit)	5/5 (100%)	0.14 to 0.17	5/5 (100%)	0.24 to 0.35	4/5 (80%)	0.05 to 0.11	5/5 (100%)	-0.09 to -0.06
	Latency next post (same subreddit)	5/5 (100%)	0.14 to 0.15	5/5 (100%)	0.24 to 0.32	3/5 (60%)	0.05 to 0.08	5/5 (100%)	-0.09 to -0.07
	Latency next comment (other subreddit)	5/5 (100%)	-0.30 to -0.28	5/5 (100%)	-0.44 to -0.37	5/5 (100%)	-0.24 to -0.20	5/5 (100%)	0.04 to 0.06

Note. "Share Significant" describes how many out of the five within-study replication analyses yielded significant results. Reported effect ranges are regression coefficients (b). The effect in the lowest tertile tests whether the slope in that area is significantly different from zero (as opposed to significantly different from the upper two thirds). The "Main Effect," is the slope of favorites (Twitter, Study 1) and score (Reddit, Study 2) z score in an LMEM without any further predictors, nonlinear or other. ^aThese models only converged with random intercepts only (no random slopes).

Table 3. Between-Person Effects Including Tweet Rate and Follower Count, in Study 1.

Sample	Latency variable	Favorites [95% CI]	Favorites—% (+1 SD)/–1 SD	Tweet rate [95% CI]	Tweet rate—% (+1 SD)/–1 SD	Favorites × Tweet rate [95% CI]	Favorites × tweet rate—% (+1 SD × 1 tweet/day)	Random effects (SE)
Filtered	Next post	0.10 [0.08, 0.12], $p < .001$	+10.73%/–9.70%	–0.65 [–0.69, –0.62], $p < .001$	+92.32%/–48.00%	0.00 [–0.01, 0.01], $p = .577$	+92.32%/–48.00%	$u_{0j} = 1.56;$ $r_{ij} = 3.51$
	Next mention	–0.16 [–0.18, –0.14], $p < .001^{\dagger}$	–14.79%/+17.35%	–0.51 [–0.56, –0.46], $p < .001$	–40.01%/+66.70%	0.02 [0.00, 0.03], $p = .001$	+2.02%/–1.98%	$u_{0j} = 2.90;$ $r_{ij} = 5.17$
	Next retweet	0.03 [0.01, 0.05], $p = .001$	+3.36%/–3.25%	–0.49 [–0.53, –0.45], $p < .001$	–38.74%/+63.23%	0.00 [–0.01, 0.01], $p = .515$	+0.30%/–0.30%	$u_{0j} = 2.34;$ $r_{ij} = 3.45$
	Next post	0.08 [0.062, 0.102], $p < .001$	+8.55%/–7.87%	–0.60 [–0.61, –0.58], $p < .001$	–45.01%/+81.85%	0.00 [–0.00, 0.02], $p = .081$	+0.80%/–0.80%	$u_{0j} = 0.83; r_{ij} = 4.16$
Unfiltered	Next mention	–0.23 [–0.25, –0.20], $p < .001^{\dagger}$	–20.23%/+25.36%	–0.47 [–0.48, –0.45], $p < .001^{\dagger}$	–37.19%/+59.20%	0.03 [0.02, 0.04], $p < .001$	+2.74%/–2.66%	$u_{0j} = 1.46; r_{ij} = 6.56$
	Next retweet	0.011 [–0.01, 0.033], $p = .300$	+1.11%/–1.09%	–0.49 [–0.51, –0.48], $p < .001$	–38.92%/63.72%	0.01 [–0.00, 0.02], $p = .276$	+0.50%/–0.50%	$u_{0j} = 1.38; r_{ij} = 4.39$
Sample	Latency variable	Favorites [95% CI]	Favorites—% (+1 SD)/–1 SD	Follower count [95% CI]	Follower count—% (doubling/halving of followers)	Favorites × follower count [95% CI]	Favorites × follower count—% (+1 SD × 1 doubling/halving)	Random effects (SE)
Filtered	Next post	0.02 [–0.07, 0.11], $p = .676$	+2.02%/–1.98%	–0.25 [–0.28, –0.22], $p < .001$	–21.81%/+27.89%	0.01 [–0.00, 0.01], $p = .072$	+0.70%/–0.70%	$u_{0j} = 2.39; r_{ij} = 3.50$
	Next mention	–0.42 [–0.54, –0.31], $p < .001^{\dagger}$	–34.36%/+52.35%	–0.19 [–0.23, –0.16], $p < .001$	–17.55%/+21.29%	0.02 [0.01, 0.03], $p < .001$	+2.22%/–2.18%	$u_{0j} = 3.38; r_{ij} = 5.18$
	Next retweet	–0.07 [–0.17, 0.02], $p = .141$	–7.04%/+7.57%	–0.19 [–0.22, –0.15], $p < .001$	–17.06%/+20.56%	0.01 [0.00, 0.02], $p = .033$	+0.80%/–0.80%	$u_{0j} = 2.78; r_{ij} = 3.45$
	Next post	0.01 [–0.09, 0.11], $p = .78$	+1.41%/–1.39%	–0.21 [–0.22, –0.20], $p < .001$	–18.78%/+23.12%	0.01 [–0.00, 0.01], $p = .129$	+0.60%/–0.60%	$u_{0j} = 1.39; r_{ij} = 4.53$
Unfiltered	Next mention	–0.44 [–0.57, –0.32], $p < .001^{\dagger}$	–35.79%/+55.74%	–0.15 [–0.16, –0.14], $p < .001^{\dagger}$	–13.84%/+16.07%	0.02 [0.01, 0.03], $p < .001$	+1.92%/–1.88%	$u_{0j} = 1.75; r_{ij} = 6.80$
	Next retweet	–0.05 [–0.16, 0.05], $p = .334$	–5.07%/+5.34%	–0.17 [–0.18, –0.16], $p < .001$	–15.63%/+18.53%	0.01 [–0.00, 0.01], $p = .197$	+0.50%/–0.50%	$u_{0j} = 1.62; r_{ij} = 4.65$

Note. Significant effects are bold-faced. † Random-intercept only model. The Follower Count was transformed with a logarithm of base 2, for simple interpretation of the effect size, as 1 unit equals a doubling or halving of the Follower Count.

Table 4. Results of the LMEMs in Study 2, With Filtered (i.e., Excluding Extremely Popular Posts) and Unfiltered (i.e., Full Data Set) Data.

Data	Latency variable	Main effect b [95% CI]	Main effect—% SD/ $\pm$ 1 SD	Lower third b [95% CI]	Lower third—% (+/ $\pm$ 1 SD)	Upper two thirds b [95% CI]	Lower third—% (+/ $\pm$ 1 SD)	Quadratic term b [95% CI]	Fixed effects (SE)	Random effects (SE)
Filtered	Next post (any subreddit)	0.16 [0.13, 0.19], $p < .001$	+17.00%/–14.53%	0.30 [0.25, 0.35], $p < .001$	+33.94%/–25.89%	0.06 [0.02, 0.10], $p = .002$	+6.45%/–6.06%	–0.08 [–0.09, –0.06], $p < .001$ [†]	$\beta_0 = 10.92 (0.05), p < .001$; $\beta_1 = 0.05 (0.02), p = .002$; $\beta_2 = 0.24 (0.03), p < .001$	$u_0 = 3.47$; $u_1 = 0.11$; $\text{Cov}(u_0, u_1) = 0.05$; $r_f = 7.21$
	Next post (same subreddit)	0.15 [0.12, 0.18], $p < .001$	+15.84%/–13.67%	0.29 [0.25, 0.33], $p < .001$	+33.65%/–25.17%	0.05 [0.02, 0.09], $p = .003$	+5.32%/–5.05%	–0.09 [–0.10, –0.07], $p < .001$	$\beta_0 = 13.27 (0.04), p < .001$; $\beta_1 = 0.05 (0.03), p = .004$; $\beta_2 = 0.24 (0.03), p < .001$	$u_0 = 2.08$; $u_1 = 0.15$; $\text{Cov}(u_0, u_1) = -0.22$; $r_f = 4.60$
	Next comment (other subreddit)	–0.29 [–0.33, –0.25], $p < .001$	–25.32%/+33.91%	–0.41 [–0.47, –0.35], $p < .001$	–33.60%/50.60%	–0.22 [–0.26, –0.17], $p < .001$	–19.35%/23.99%	0.06 [0.03, 0.08], $p < .001$ [†]	$\beta_0 = 12.16 (0.08), p < .001$; $\beta_1 = -0.22 (0.03), p < .001$; $\beta_2 = -0.19 (0.04), p < .001$	$u_0 = 2.85$; $u_1 = 0.02$; $\text{Cov}(u_0, u_1) = -0.14$; $r_f = 3.36$
Unfiltered	Next post (any subreddit)	0.08 [0.06, 0.11], $p < .001$	+8.65%/–7.96%	0.32 [0.25, 0.40], $p < .001$	+38.23%/–27.65%	0.05 [0.02, 0.07], $p < .001$	+4.90%/–4.67%	–0.01 [–0.02, –0.00], $p = .003$ [†]	$\beta_0 = 10.90 (0.04), p < .001$; $\beta_1 = 0.08 (0.01), p < .001$; $\beta_2 = 0.27 (0.04), p < .001$	$u_0 = 3.48$; $u_1 = 0.5$; $\text{Cov}(u_0, u_1) = 0.06$; $r_f = 7.01$
	Next post (same subreddit)	0.07 [0.06, 0.09], $p < .001$	+7.47%/–6.95%	0.33 [0.27, 0.39], $p < .001$	+39.39%/–28.26%	0.03 [0.01, 0.05], $p < .001$	+3.25%/–3.15%	–0.02 [–0.03, –0.01], $p < .001$	$\beta_0 = 13.24 (0.03), p < .001$; $\beta_1 = 0.07 (0.008), p < .001$; $\beta_2 = 0.30 (0.03), p < .001$	$u_0 = 2.00$; $u_1 = 0.01$; $\text{Cov}(u_0, u_1) = -0.32$; $r_f = 4.21$
	Next comment (other subreddit)	–0.35 [–0.38, –0.32], $p < .001$	–29.46%/+41.76%	–1.14 [–1.24, –1.04], $p < .001$	–68.04%/212.90%	–0.23 [–0.27, –0.20], $p < .001$	–20.90%/26.42%	0.06 [0.05, 0.07], $p < .001$ [†]	$\beta_0 = 11.80 (0.08), p < .001$; $\beta_1 = -0.23 (0.02), p < .001$; $\beta_2 = -0.91 (0.05), p < .001$	$u_0 = 11.97$; $u_1 = 0.16$; $\text{Cov}(u_0, u_1) = 0.11$; $r_f = 9.12$

Note. Significant effects are bold-faced. [†] Random-intercept only model. Parameters: intercept (β_0), predictor score (β_1), discretizing factor (β_2), intercept variance (u_0), slope variance (u_1), covariance (u_0, u_1), residual variance (r_f).

as their profile name). We sampled roughly 10,000 users until 2,000 users met the criteria. For the full list of criteria, see OSF. The final data set consisted of 47,399 posts from 2,000 users and posted between August 2012 and November 2022. Data were randomly split into five random sets of 1,000 users to replicate the findings.

Measures. Please see OSF for the exact calculation of the different variables.

Favorite Count. The favorite count was Anscombe-transformed and standardized within individuals (i.e., centering within cluster) to produce individual-level estimates (Enders & Tofighi, 2007).

Latency Until Next Post. Latency until the next post indicates the time that passes until a user posts again, following their last post. For *threads* (i.e., a coherent string of posts), we only analyzed the first post to avoid misinterpretation of artificially low latencies until the next post in the thread.

Latency Until Next Mention. Latency until the next mention indicates the time that passes until a user posts a mention again, following their last post.

Latency Until Next Retweet. Latency until the next retweet indicates the time that passes until a user retweets a post, following their own last post.

Results

Sample Characteristics. The median of users' favorite count was 15.00 ($M = 58.43$), and the median user posted 1.39 times a week ($M = 5.06$). Information on users' demographics (e.g., age, gender, and residence) is not available.

Latency Until the Next Post. We found a significant positive relationship between the popularity of the last post and the latency until the next post (Table 1). We introduced a discretizing factor, dividing the slope estimates into the lowest third of popularity (based on cluster-standardized measurements) and the upper two thirds representing more popular posts. We observe steeper positive slopes in the lowest third compared with the two upper thirds (Table 1).

Within-study replication confirmed a significant positive relationship between post popularity and latency until the next post, as well as steeper positive slopes in the lowest third compared with the two upper thirds (Table 2). Incorporation of this discretizing factor resulted in a significantly improved model fit, $\chi(1)^2 = 18.95$, $p < .001$. Our findings thus suggest a nonlinear relationship between the popularity of a post and the latency until the next post, with a steeper relationship in the lower third of the distribution (Figure 1). This suggests that users post more quickly following an unusually unpopular post, rather than posting

particularly slowly following a very popular post. Findings thus align with the Reconnection with Many Hypothesis and suggest no evidence for the Withdrawal Hypothesis.

Latency Until the Next Mention. The Reconnection with Specific Individuals Hypothesis predicts a *positive* association of relative post popularity and latency until the next mention. However, we found a significant *negative* relationship between the popularity of the last post and the latency until the next mention (Table 1). Compared with the upper two thirds, the slope was less steep in the lowest third (Table 1). This suggests that the negative relationship between the popularity of the previous post and the latency until the next mention is attributable to particularly quick mentions following *more* popular posts (Figure 1).

Within-study replication also shows a significant negative relationship between post popularity and latency until the next mention, which was also reflected in the upper two thirds of popular posts. However, when considering only the lower third of popularity, the relationship between post popularity and mentioning latency was not significantly different from zero (Table 1). The discretizing factor significantly improved model fit, $\chi(1)^2 = 18.95$, $p < .001$. Our findings suggest a nonlinear relationship between the popularity of a post and the latency until the next mention, with the strongest relationship in the upper two thirds of the distribution. Findings are thus in disagreement with the Reconnection with Specific Individuals Hypothesis.

Latency Until the Next Retweet. We found a significant relationship between the popularity of the last post and latency until the next retweet (Table 1). The slope was markedly steeper in the lowest third than the slope in the upper two thirds, for which there was no significant relationship between popularity and latency until the next retweet (Table 1). Within-study replication also showed consistent patterns (Table 2). Introducing the discretizing factor improved model fit, $\chi(1)^2 = 8.09$, $p = .004$. Taken together, there was a weak but general trend that users retweeted more quickly after less popular posts, in line with the Reconnection with Specific Individuals Hypothesis.

Exploratory Analyses Including Extremely Popular Posts. In exploratory fashion, we repeated all analyses with the full data set, that is, not excluding extremely popular posts to gauge generalizability beyond examining unpopular posts. The results were generally highly robust and not affected by including highly popular posts (Table 1).

Exploratory Within-Study Replication Analyses. To replicate the results, we used a random subsampling method (Molinaro et al., 2005), drawing five random samples of $N = 1,000$ from the total sample to assess the results' consistency across subsamples. The vast majority of relationships were robust (Table 2). Given that the quadratic continuous predictor

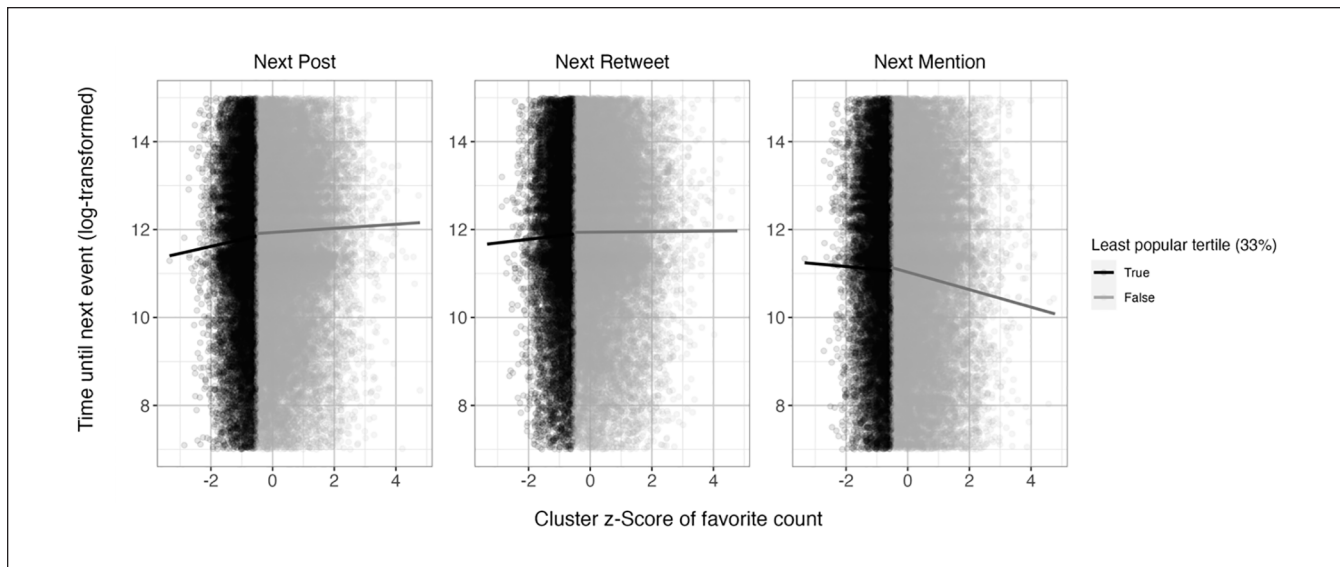


Figure 1. Association of Relative Favorite Count of a Post With the Time Taken Until the Next Own Tweet (Left Plot), Mention (Plot in the Middle), and Retweet (Right Plot), in Study 1.

Note. Separate slopes are calculated for the least popular third of all posts versus the remaining posts.

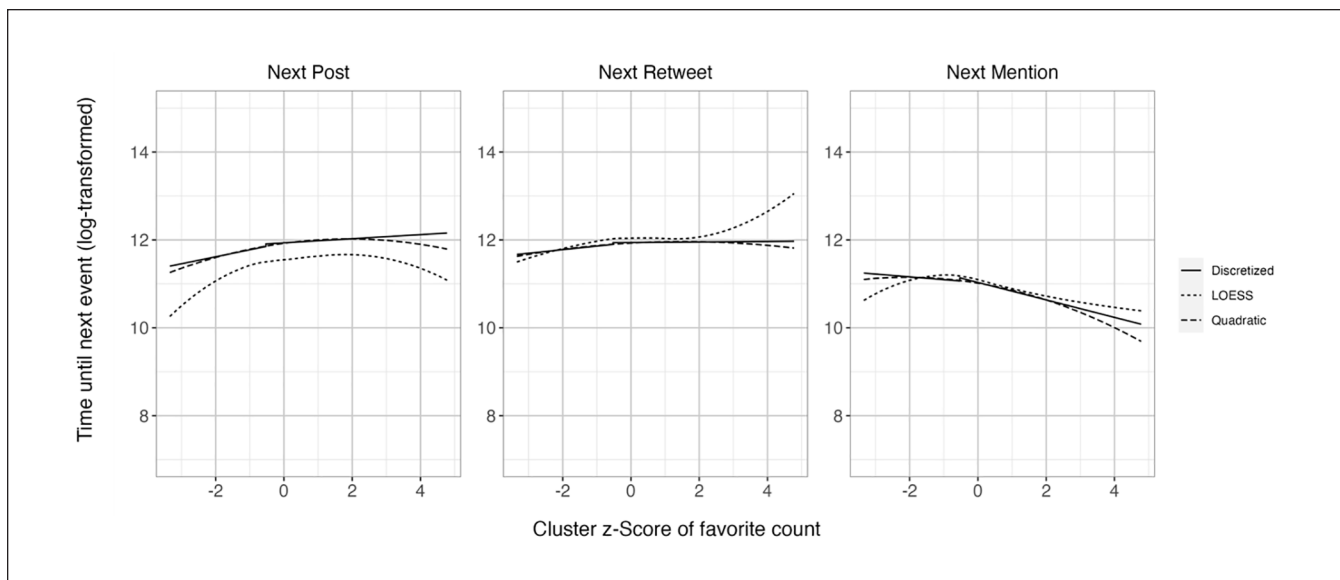


Figure 2. Association of Relative Favorite Count of a Post With the Time Taken Until the Next Own Tweet (Left Plot), Mention (Plot in the Middle), and Retweet (Right Plot), in Study 1, Comparing Discretized (Tertile Split), Quadratic Model, and LOESS Curves.

proved to be a good fit (see OSF), Figure 2 shows the discretized tertile split, a model with continuous quadratic predictor, and the LOESS (locally estimated scatterplot smoothing) curve, a regression curve that is estimated within smaller subsets and can thus display nonlinear functions without a priori assumptions (Cleveland, 1979). The quadratic term was significant for almost all analyses (Table 1), indicating nonlinear relationships between the favorite count and the latency until the next post, the next mention (only in the filtered data set), and the next retweet. All quadratic

coefficients are negative. For latency until the next post and for latency until the next retweet, this means that both after very unpopular posts and after very popular posts, users post and retweet again more quickly. For the latency until the next mention, the negative quadratic coefficient indicates that both after very unpopular posts and after very popular posts, users mention others again more quickly.

Between-Person Effects of Follower Count and Tweet Rate. As preregistered, we also tested whether the average tweet rate

of the user (i.e., how often a user posted per day, only considering own content) and/or the number of followers of the user moderates the tested relationships. Analyses of tweet rate indicated that for the latency until the next mention, the effect of favorites was stronger for users with a lower tweet rate (Table 3). Analyses of follower count indicated that for the latency until the next mention and until the next retweet (only filtered data set), the effect of favorites was stronger for users with fewer followers (Table 3).

Discussion

In Study 1, we investigated user behavior on Twitter, examining the relationship between the popularity of posts and time until the next user activity. One striking observation is that users tend to post more quickly following an unpopular tweet, potentially as a way to regain lost connection in line with the Reconnection with Many Hypothesis and contradicting the Withdrawal Hypothesis. As an alternative interpretation of the observed effects, users may post again more quickly in an attempt to influence the Twitter algorithm (i.e., if one post receives many favorites, another post is also more likely to be displayed in other users' feeds because it is similar content). Put differently, posting again quickly may be a strategic way to increase exposure to the initial Tweet. We address this limitation with Study 2 because on Reddit, posts are displayed chronologically in specific subreddits, and therefore, posting again quickly does not influence the algorithm in the same way as it would on Twitter.

We observed a weaker, but comparable trend in the latency until the next retweet: Users retweet more quickly the less popular their previous post was, in line with the Reconnection with Specific Individuals Hypothesis. However, the strength of this relationship was not clearly intensified in the lower third of the distribution, suggesting that retweet latency is either less impacted by having an unpopular tweet, or more impacted by having a popular tweet.

Interestingly, when it came to the latency of mentions, the pattern reversed: There was a negative relationship between the popularity of a previous post and the latency until the next mention, meaning that individuals mentioned others more quickly after more *popular* posts. This trend was weaker in the lowest third of the popularity distribution, which differed from the patterns observed for posting and retweeting. Thus, the observed mentioning pattern does not align with the Reconnection with Specific Individuals Hypothesis. However, we caution against drawing definitive conclusions from this finding because the positive relationship could be a result of decreased time until the next activity in response to one's own more popular post. That is, after having a particularly popular post, many others may comment on that post, resulting in more mentions when the users interact with those comments (e.g., "@user, thank you for your comment!"). Therefore, quicker mentioning activity

after a popular post does not necessarily reflect an active engagement with other users; instead, the engagement could be reactive, in line with reciprocity norms (Gouldner, 1960). We aim to delineate this finding further in Study 2 by focusing on commenting latency that is unrelated to one's own post.

We also tested for quadratic relationships between favorite count and latency until the next behavior, finding support for quadratic relationships for all latency variables. This indicates that users post, retweet, and mention others more quickly after very unpopular and after very popular posts. This observation aligns with both our Reconnection Hypotheses and with previous findings that individuals use a reward-learning approach to using social media: Getting many favorites is rewarding and may prompt individuals to use social media more to obtain more psychological reward (Lindström et al., 2021). Importantly, the results were largely consistent when extremely popular posts were not filtered out, speaking in favor of the generalizability of the results.

Moreover, we found that users who posted less frequently and had fewer followers reacted more strongly to having an unpopular tweet in that they posted, mentioned, and retweeted others more quickly. Put differently, those who post more frequently and those who have more followers seem to be less affected by having one unpopular post. This could mean that the emotional impact of a single post's performance is lower for frequent and for posters with a big followership (e.g., E. Lee, Lee, et al., 2015). Moreover, if follower count can be interpreted as an indicator of social power on Twitter, our findings align with previously found buffering effects of higher social power against the negative effects of social exclusion (Kuehn et al., 2015). However, the between-person effects of tweet rate and follower count were not consistent across the dependent variables and, therefore, have to be interpreted cautiously.

In summary, rather than observing a general increase or decrease in time until the next activity following an unpopular post, we found that the time until more public activities (posting and retweeting, which reach all of one's followers) tended to increase, while the time until activities aimed at specific others (mentions, which directly address only one user) tended to decrease. This might indicate that users are looking to rebound from the lack of crowd affirmation (reflected in the number of favorites) following an unpopular post, in line with the Reconnection with Many Hypothesis, as opposed to seeking validation from individual users, in what would be in line with the Reconnection with Specific Individuals Hypothesis.

Study 2

Study 2 extends Study 1 in multiple ways by examining another context, Reddit. Reddit users may post links to articles, videos, or other content, and other users can vote on the content's popularity. Reddit is organized in *subreddits* that

are focused on specific topics. This offers the chance to look at commenting latency in *other* subreddits, that is, posting comments that are not directly related to one's own content and addressing a major limitation of Study 1. Another distinction is Reddit's emphasis on user anonymity. While both Twitter and Reddit can be used anonymously in theory, many users choose a Twitter profile name that makes them identifiable to some extent (e.g., Peddinti et al., 2014). Reddit is one of the most anonymous platforms at the time of this contribution (e.g., Triggs et al., 2021). Anonymity may play a key role in how individuals use social media: Users with a higher degree of anonymity post more (e.g., Peddinti et al., 2014) and more honest content (e.g., Karahanna et al., 2018). Thus, identifying whether behavior following unpopular posts differs based on users' anonymity was another important endeavor of Study 2.

Users can upvote or downvote content on Reddit. To achieve comparability with the relative favorite count on Twitter used in Study 1, we used the sum of upvotes and downvotes (i.e., *score*) that is displayed in Reddit as the predictor variable in this study.

We test the following preregistered hypotheses in Study 2:

Hypothesis 1: "Reconnection with Many": Submissions with a lower (vs higher) within-person-centered score have smaller latency until the next submission (a) in any subreddit and (b) in the same subreddit (i.e., the one of the previous submission).

Hypothesis 2: "Reconnection with Specific Individuals": Submissions with a lower (vs higher) within-person-centered score have smaller latency until the next comment in another subreddit.³

Again, we also preregistered the assumption that the effects described in Hypotheses 1 and 2 will be significantly stronger for submissions at the lower end of the distribution of scores, indicating that effects are driven by unpopular rather than average or very popular submissions. For Study 2, we explicitly preregistered splitting the distribution into the lower third and the two upper thirds of the distribution.

Given the evidence against the Withdrawal Hypothesis (Hypothesis 3) in Study 1, we did not preregister the Withdrawal Hypothesis in Study 2.

Method

Social Media Data. For Study 2, we obtained data from 2,000 Reddit users in 10 different subreddits (i.e., thematic collections of posts). We chose the 10 subreddits with the most followers at the time of preregistration (i.e., March 2023) that had at least 200 users who contributed 10 or more posts (i.e., "funny," "aww," "todayilearned," "pics," "Show-erthoughts," "food," "EarthPorn," "Art," "gifts," and "Old-SchoolCool") that mainly contained individual content

creator's own content (e.g., no help requests or sharing of official content—for a more detailed list, see the subreddit selection script via OSF Supplement 2c). For each of the subreddits, we sampled frequent contributors to this subreddit (≥ 10 contributions). For each user, we exclusively analyzed posts from the subreddit where they were frequently active. As our dependent variables, we used latencies until users' next post, considering posts within the same subreddit and those in other subreddits, and we used latency until the next comment in another subreddit. The final data set consisted of 58,442 posts from 2,000 users and posted between November 2008 and May 2023.

Measures. The measures in Study 2 mirrored the measures in Study 1, with minor differences due to differences between Twitter and Reddit.

Score. Different from Twitter, Reddit users can choose to either upvote or downvote a post. However, on Reddit, a post's visibility is disproportionately influenced by its upvote count (see https://www.reddit.com/r/blog/comments/o5tjcn/evolving_the_best_sort_for_reddits_home_feed/?rdt=50386). Posts with many downvotes become less visible, as the algorithm prioritizes posts with more upvotes. Posts receive downvotes only when they are still visible, that is, when they are popular. Consequently, a high number of downvotes are often an indicator of a post's reach and visibility. Therefore, we used the sum score of upvotes and downvotes that is also displayed on Reddit. Note that all scores in the data set were positive. The score was transformed in the same fashion as the favorite count in Study 1.

Latency Until Next Post. Due to Reddit's subreddit structure, we can differentiate between users' behavior in the same subreddit versus in other subreddits. We, therefore, differentiate between latency until the next post *in any subreddit* and *in the same subreddit as the last post*.

Latency Until Next Comment (Other Subreddit). We analyze latency of the next comment *in any other subreddit*, that is, other than the one where the respective post was posted. This allows us to control for shorter latencies of commenting due to the popularity of a previous post, which we could not control for in the Twitter data. Comments in Reddit are similar to mentions in Twitter, as both signify direct responses to another person's post.

Results

Sample Characteristics. Users received a median score of 56.98 ($M = 513.77$) on their posts in the subreddit they were sampled from and the median user posted 0.98 times a week ($M = 2.48$) in any subreddit. Users posted 15.37% of their posts in the subreddit they were sampled from on average (for further descriptive analyses see OSF Supplement 2a).

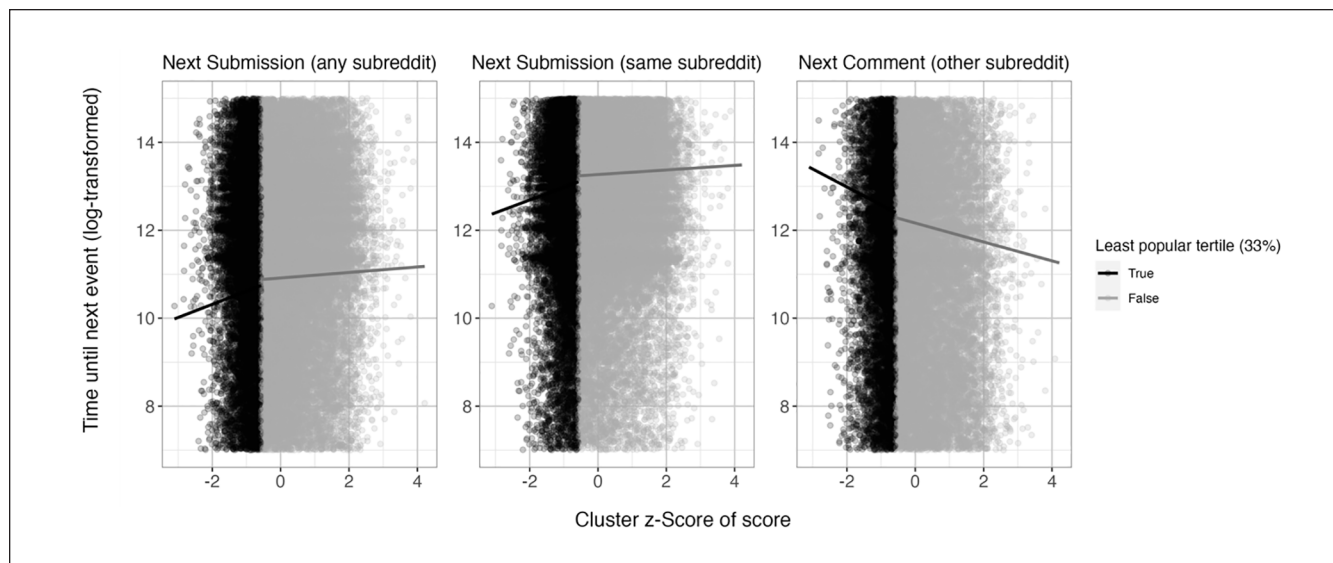


Figure 3. Association of Relative Score of a Post With the Time Taken Until the Next Post in Any Subreddit (Left Plot), the Next Post in the Same Subreddit (Plot in the Middle), and the Next Comment in Another Subreddit Than That of the Previous Post (Right Plot), in Study 2.

Note. The slopes in the graphs are based on the regression coefficients of the linear mixed models. Separate slopes are calculated for the least popular third of all posts versus the remaining posts.

Latency Until the Next Post

In Any Subreddit. We found a significant positive relationship between the popularity of the last post and latency until the next post in any subreddit. We again introduced a discretizing factor, dividing the slope estimates into the lowest third of popularity and the upper two thirds representing the more popular posts. The association was significant for both slopes, but we observed a steeper slope in the lowest third (Table 4). The discretizing factor improved model fit, $\chi(1)^2 = 50.55, p < .001$. Our findings suggest a nonlinear relationship between the score of a post and the latency until the next post in any subreddit, with the strongest relationship appearing in the lower third of the distribution (Figure 3). Thus, users tend to post more quickly following an unusually unpopular post, in line with the Reconnection with Many Hypothesis.

In the Same Subreddit. Parallel findings emerged for the latency until the next post in the *same* subreddit. We observed a significant positive relationship between popularity of the last post and the latency until the next subreddit-specific post. The discretizing factor revealed similar patterns: The association was significant for both slopes, but the slope was steeper in the lowest third (Table 4 and Figure 3). The discretizing factor again improved model fit, $\chi(1)^2 = 77.44, p < .001$, mirroring the findings from the post latency in any subreddit, in line with the Reconnection with Many Hypothesis.

Latency Until the Next Comment (Other Subreddits). Regarding latency until the next comment in other subreddits⁴ (i.e.,

other than that of the previous post), we found a significant *negative* relationship between the popularity of the last post and the latency until the next comment. This suggests that after posting less popular content in a specific subreddit, users tend to engage in other communities less quickly. Again, the association was significant for both slopes, but the slope was steeper in the lowest third (Table 4). The discretizing factor improved model fit, $\chi(1)^2 = 23.18, p < .001$. Our findings suggest a nonlinear relationship between the popularity of a post and the latency until the next comment in a different subreddit, with the strongest relationship appearing in the lower third of the distribution (Figure 3). Thus, users tend to comment more slowly following an unusually unpopular post, in disagreement with the Reconnection with Single Individuals Hypothesis.

Exploratory Analyses Including Extremely Popular Posts. Again, we repeated all analyses with the full data set, not excluding extremely popular posts. Again, the pattern of results was highly robust (Table 4).

Exploratory Within-Study Replication Analyses. Again, to replicate the results, we used a random subsampling method (Molinari et al., 2005), finding that the vast majority of relationships were robust (Table 2). Again, given that the quadratic continuous predictor proved to be a good fit (see OSF), Figure 4 shows the discretized tertile split, a model with continuous quadratic predictor, and the LOESS curve. As in Study 1, the quadratic term was significant for all analyses (Table 4), indicating nonlinear relationships between Reddit

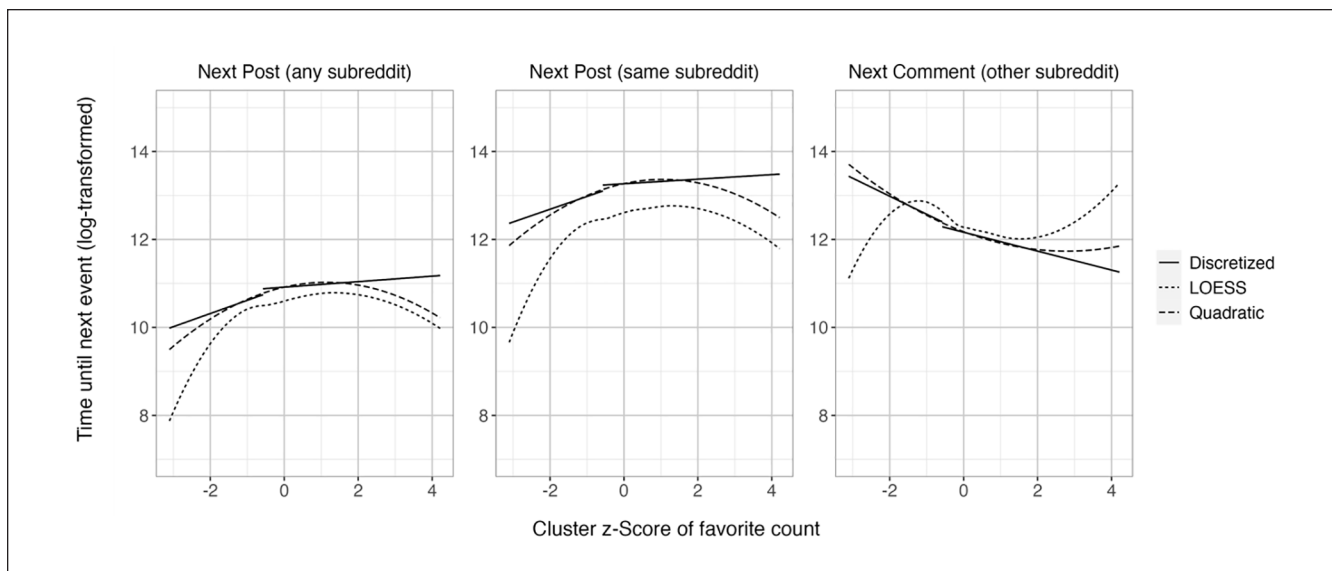


Figure 4. Association of Relative Score of a Submission With the Time Taken Until the Next Post in Any Subreddit (Left Plot), Next Post in the Same Subreddit (Plot in the Middle), and Next Comment in Another Subreddit (Right Plot), in Study 2, Comparing Discretized (Tertile Split), Quadratic Model, and LOESS Curves.

Note. Even though the LOESS curve for comments may look cubic, LOESS plots that rely on very small samples in the margins are subject to a lot of uncertainty. Indeed, the cubic model is a worse fit than the quadratic and the discretized model, see OSF.

scores and all latency variables. For the latency until the next submission in any and the same subreddit, the quadratic coefficients were negative, meaning that after very unpopular submissions and after very popular submissions, users post again more quickly in any and the same subreddit. For the latency until the next comment, the quadratic coefficient was *positive*, meaning that after very unpopular submissions and after very popular submissions, users comment in other subreddits more slowly.

Between-Person Effects of Submission Rate. As preregistered, we also tested whether the average submission rate (i.e., how often a user posted per day, only considering own content) moderates the tested relationships. Indeed, the effects of score were stronger for users with a lower submission rate for next submission in any subreddit (only filtered data set) and for next comment in another subreddit (Table 5).

Discussion

Study 2 sought to corroborate Study 1's findings on the relationship between the popularity of posts and time until the next user activity on another platform: Reddit. We found that users tend to post more quickly following an unusually unpopular post, both in any subreddit and in the same subreddit, in line with the Reconnection with Many Hypothesis.

By focusing on comments that are unrelated to one's own post, Study 2 delineated quicker commenting activity as an artifact of responses to an own popular post. In this study, individuals took a particularly long time before commenting

in other subreddits, following highly unpopular posts. Study 2 thus overall aligns with observations from Study 1 in that there is not a general increase or decrease in the time until the next activity following an unpopular post. Rather, activities aimed at reconnecting with many others, such as posting, are performed faster, while activities targeting specific others, such as commenting, are performed more slowly, lending further support to the Reconnection with Many Hypothesis over the Reconnection with Specific Individuals Hypothesis. Again in line with Study 1, all results could be replicated when extremely popular posts were not filtered out. Furthermore, we again found support for quadratic relationships between Reddit scores and all latency variables: After very unpopular submissions and after very popular submissions, users posted again more quickly but commented less quickly. This observation again aligns with the Reconnection with Many Hypothesis and with social media user's reward-learning approaches after popular posts (Lindström et al., 2021). Moreover, Study 2 also documents that those who post more frequently on Reddit are less affected by having one unpopular post. Interestingly, less frequent posters only posted more quickly in any subreddit, not the same subreddit, which could indicate that they sought the lacking affirmation in a different community.

General Discussion

Social media and experiences of social exclusion are part of the reality of people's everyday lives around the globe (e.g., Büttner, Ren, et al., 2024; *Forbes*, 2023). We investigated

Table 5. Between-Person Effects Including Submission Rate, in Study 2.

Sample	Latency variable	Score [95% CI]	Score—% (+1 SD/–1 SD)	Submission rate [95% CI]	Submission rate—% (+1 SD/–1 SD)	Score × submission rate [95% CI]	Score × submission rate—% (+1 SD × 1 submission/day)	Random effects (SE)
Filtered	Next submission (any subreddit)	0.17 [0.14, 0.20], <i>p</i> < .001	+18.65%/–15.72%	–0.63 [–0.71, –0.55], <i>p</i> < .001	–46.58%/+87.19%	–0.04 [–0.07, –0.01], <i>p</i> = .007	–3.82%/+3.98%	<i>u</i> _{0j} = 3.05; <i>r</i> _{ij} = 7.21
	Next submission (same subreddit)	0.15 [0.12, 0.18], <i>p</i> < .001 [†]	+15.72%/–13.58%	–0.35 [–0.41, –0.29], <i>p</i> < .001	–29.60%/+42.05%	0 [–0.03, 0.03], <i>p</i> = 1	0%/0%	<i>u</i> _{0j} = 1.95; <i>r</i> _{ij} = 4.61
	Next comment (other subreddit)	–0.31 [–0.35, –0.27], <i>p</i> < .001	–26.51%/+36.07%	–0.36 [–0.52, –0.20], <i>p</i> < .001	–30.23%/+43.33%	0.05 [0.01, 0.08], <i>p</i> = .022	+4.60%/–4.40%	<i>u</i> _{0j} = 11.38; <i>r</i> _{ij} = 8.78
	Next submission (any subreddit)	0.09 [0.06, 0.11], <i>p</i> < .001	+9.20%/–8.42%	–0.64 [–0.72, –0.56], <i>p</i> < .001	–47.22%/+89.46%	–0.01 [–0.04, 0.01], <i>p</i> = .234	–1.39%/+1.41%	<i>u</i> _{0j} = 3.04; <i>r</i> _{ij} = 7.01
Unfiltered	Next submission (same subreddit)	0.07 [0.05, 0.09], <i>p</i> < .001 [†]	+7.36%/–6.85%	–0.35 [–0.41, –0.29], <i>p</i> < .001 [†]	–29.67%/+42.19%	–0.00 [–0.02, 0.02], <i>p</i> = .894	–0.10%/+0.10%	<i>u</i> _{0j} = 1.86; <i>r</i> _{ij} = 4.21
	Next comment (other subreddit)	–0.37 [–0.40, –0.34], <i>p</i> < .001	–30.86%/+44.63%	–0.32 [–0.48, –0.16], <i>p</i> < .001	–27.60%/+38.13%	0.06 [0.02, 0.09], <i>p</i> = .001	+5.97%/–5.64%	<i>u</i> _{0j} = 11.73; <i>r</i> _{ij} = 9.16

Note. Significant effects are bold-faced. [†] Random-intercept only model.

how users react to social exclusion online, using observational data from two large social media platforms, Twitter and Reddit. Both studies show that having a relatively unpopular post on social media motivates users to post again more quickly to a broad audience, in line with what we term the Reconnection with Many Hypothesis. In Study 1, this means that Twitter users who just experienced having a relatively unpopular tweet (i.e., by one standard deviation, in the lowest third of popularity), post again more quickly by 14.69% of the average time until the next post (Table 1). In Study 2, the effects were even stronger: For instance, a Reddit user who just experienced having a relatively unpopular Reddit post (i.e., by one standard deviation, in the lowest third of popularity) posts again more quickly by 25.89% of the average time until the next submission in any subreddit, and more quickly by 25.17% of the average time until the next submission in the same subreddit (Table 4). However, exclusion does not motivate users to target specific others, such as via comments or mentions.

Our finding that there is no general increase in the time until any next activity following a relatively unpopular post may be interpreted as a form of acknowledgment-seeking that is specific to experiencing exclusion on social media. Rather than a general need to reconnect with others, be that, a broad audience or specific others, exclusion on social media seems to specifically motivate acknowledgment-seeking actions like posting content again quickly. In line with the Reconnection with Many Hypothesis, our findings speak in favor of individuals behaving to be *acknowledged* rather than to be *liked* after social media exclusion. By posting more quickly after an unpopular post, social media users may regain acknowledgment from others (Karahanna et al., 2018) and re-plenish needs that are threatened by exclusion (e.g., Lutz et al., 2022; Williams, 2009). Furthermore, as hypothesized, the effects were stronger in the least popular third of posts, suggesting that the effect was driven by the difference in responses following unpopular versus regular posts, rather than regular versus popular posts.

Reversely, individuals seem to refrain from targeting specific others to reconnect, such as via comments or mentions, lending no support to the Reconnection with Specific Individuals Hypothesis. Potentially, targeting specific others entails a higher social risk of being excluded again because specific individuals might (intentionally or unintentionally) ignore a comment or mention. This aligns with findings that targets of exclusion are hesitant to share their experiences with others because they fear further exclusion (e.g., Meral et al., 2021). Moreover, even though the degree of a user's anonymity may influence their social media use (e.g., Karahanna et al., 2018; Peddinti et al., 2014), we observed that it does not matter whether the excluded users were completely anonymous (Reddit) or not (Twitter).

We here focus on the *relative* unpopularity of users' content as an indication of social exclusion, comparing the number of paralinguistic digital affordances (here: favorites on

Twitter and upvote-downvote-scores on Reddit, Hayes et al., 2016) that users received to the number of paralinguistic digital affordances they usually received. We, therefore, focused on having a relatively unpopular post, rather than general behavior on social media. Previous experimental research has extensively shown that receiving fewer paralinguistic digital affordances (e.g., Likes) than others or receiving no Likes at all is perceived as exclusionary and threatens psychological needs that are typically threatened after exclusion (e.g., Lutz & Schneider, 2020; Reich et al., 2018). We extend this approach by using the popularity of users' posts in light of to the *usual* popularity of their posts, that is, *relative* popularity. Relative popularity might be more informative than the absolute number of Likes since individuals who usually receive few Likes might experience psychological benefits even from receiving one or two Likes. This idea aligns with the concept of "Power of One" holding that being included even by one person alleviates the sting of exclusion (e.g., Büttner, Jauch, et al., 2024; DeWall et al., 2010).

Although not the focus of this contribution, we also find that users with very popular posts engage more quickly in acknowledgment-seeking behaviors. This aligns with recent work by Lindström et al. (2021) showing that individuals approach social media with a reward-learning mindset so that more popular posts lead to faster posting.

The present findings rely on individuals who post social media content with a minimum frequency. As such, these findings may generalize better to individuals who post regularly as opposed to those who post rather infrequently. Still, we believe that our findings generalize to any social media user who regularly posts content. We focused on two mainly text-based platforms, Twitter and Reddit; however, image-based platforms such as Instagram or Facebook may carry more emotionality and may be more intimate than text (e.g., E. Lee, Lee, et al., 2015; Pittman & Reich, 2016). Social media research has to be mindful of the ephemerality of single platforms and their features when making statements about generalizability (see Büttner et al., 2023 for a discussion). We believe that our findings generalize to other, also future, social media platforms, as long as they entail user-generated posted content and have mechanisms of paralinguistic affordance that signal approval (e.g., Hayes et al., 2016).

The present findings are silent about the effectiveness of different behaviors following exclusion on social media. The effectiveness of posting again quickly after exclusion on social media for recovering threatened needs may depend on multiple factors, such as who reacts to the post, how quickly other social media users react, but also, trait-level moderators of the excluded user, such as mental health factors (e.g., Büttner et al., 2021), or chronic experiences of exclusion (Büttner, Jauch, et al., 2024, Study 2). Future research could combine individuals' social media data (e.g., Zhao et al., 2023) with experiments (e.g., Lindström et al., 2021) to examine users' reactions to exclusion and effects of social media activity on recovery.

Another avenue for future research is exploring the content of unpopular posts. The present findings are silent about the degree of psychological threat that each exclusion on social media conveys. How hurtful exclusion is and how individuals behave afterward might, at least partly, depend on the emotionality of the unpopular post and the level of need threat elicited by the exclusion experience (Büttner, Ren, et al., 2024). Just as being briefly ignored in an elevator may be easier to recover from than being ignored by a loved one for extended periods of time (e.g., Williams, 2009), having an emotionally important social media post ignored might hurt more and have other effects on behavior. Relatedly, future research should examine under which circumstances individuals turn to social media following exclusion experiences in real life.

Future research may use text analyses to probe for the degree of pro- or antisociality of posts and comments following social exclusion. Exclusion may drive antisocial behavior such as aggression or violence, mainly because antisocial behavior alleviates threats to power and provocation needs by provoking acknowledgment (e.g., Ren et al., 2018). If social media exclusion indeed mainly triggers acknowledgment-seeking behavior, then social media exclusion may have important implications for antisocial behavior such as trolling, hate speech, or cyberbullying (e.g., Castaño-Pulgarín et al., 2021; Yokotani & Takano, 2021).

Feeling excluded on social media and posting again more quickly could lead to a downward spiral of not getting the attention that one craves and posting again more quickly to be acknowledged by others. Specifically, if a user's post received comparatively little attention and they post faster as a way of coping, that post could receive comparatively little attention as well because the post is qualitatively inferior, or, because other users may feel bothered by someone who posts a lot of content quickly (i.e., spammers, e.g., Kang et al., 2013). This spiral may foster social media addiction over time (e.g., Poon, 2018).

Conclusion

Experiences of social media exclusion decrease the time until users post again, likely as a form of acknowledgment-seeking. Conversely, the time until efforts are taken to reconnect with specific others increases. Using large-scale social media data from two different platforms, we provide a broad perspective on online exclusion's behavioral consequences, contributing to our understanding psychological implications of online communication dynamics.

Acknowledgments

The authors thank Daan Brückner-Collet for help with preparing the analysis code.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by grant RU 2359/4-1 from the German Research Federation (DFG) awarded to Selma C. Rudert and grant #P000PS_207236/1 from the Swiss National Science Foundation (SNSF) awarded to Christiane M. Büttner.

Open Science Statement

All data, analysis code, and supplemental analyses are available via: https://osf.io/h62gz/?view_only=13360594eb654ab4a690ebdd11a4da00.

ORCID iDs

Christoph Kenntemich  <https://orcid.org/0000-0002-9751-0773>

Christiane M. Büttner  <https://orcid.org/0000-0001-6296-2756>

Selma C. Rudert  <https://orcid.org/0000-0001-5986-2447>

Supplemental Material

Supplemental material is available online with this article.

Notes

1. Robustness tests (see OSF Supplement 2a, l. 618ff and Supplementary robustness tests table) indicate that the pattern of results in Study 2 remains the same when mimicking exclusion of extremely popular posts in Study 1 and excluding only the 13% most popular Reddit posts.
2. At the time of data collection, user verification (blue ticks) was not available for purchase yet.
3. We also preregistered additional analyses of the count of comments in the 24 hr and the potential moderating effect of submission rate. These analyses are reported as supplementary: OSF.
4. We preregistered to analyze latency until the next comment in any subreddit; however, controlling for posts in the same subreddit allows us to ensure that decreased latencies do not originate from the popularity of a post itself (e.g., engaging with others under one's own popular post). Analyses with any subreddit and same subreddit can be found online (OSF).

References

- Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3–4), 246–254. <https://doi.org/10.1093/biomet/35.3-4.246>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, 2(4), 396–403. <https://doi.org/10.1111/j.1745-6916.2007.00051.x>
- Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker,

- B. M. (2017). GlmmTMB balances speed and flexibility among packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, 9(2), 378–400. <https://doi.org/10.32614/RJ-2017-066>
- Büttner, C. M., & Greifeneder, R. (2024). Everyday ostracism experiences of depressed individuals: Uncovering the role of attributions using experience sampling. *Journal of Affective Disorders Reports*, 17, 100804. <https://doi.org/10.1016/j.jadr.2024.100804>
- Büttner, C. M., Jauch, M., Marinucci, M., Williams, K. D., Greifeneder, R., Riva, P., & Rudert, S. C. (2024). It will (never) stop hurting: Do repeated or chronic experiences of exclusion lead to hyper- or hyposensitive psychological responses? *Group Processes & Intergroup Relations*, 27(2), 256–277. <https://doi.org/10.1177/13684302221140002>
- Büttner, C. M., Lalot, F., & Rudert, S. C. (2023). Showing with whom I belong: The desire to belong publicly on social media. *Computers in Human Behavior*, 139, 107535. <https://doi.org/10.1016/j.chb.2022.107535>
- Büttner, C. M., Ren, D., Stavrova, O., Rudert, S. C., Williams, K. D., & Greifeneder, R. (2024). Ostracism in everyday life: A framework of threat and behavioral responses in real life. *Journal of Personality and Social Psychology*. Advance Online Publication. <https://doi.org/10.1037/pspi0000471>
- 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, 107062. <https://doi.org/10.1016/j.chb.2021.107062>
- Büttner, C. M., Rudert, S. C., & Greifeneder, R. (2021). Depressed and excluded: Do depressive symptoms moderate recovery from ostracism? *Journal of Affective Disorders*, 294, 730–736. <https://doi.org/10.1016/j.jad.2021.07.075>
- Castañó-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, 58, 101608. <https://doi.org/10.1016/j.avb.2021.101608>
- Chen, Z., Poon, K.-T., DeWall, C. N., & Jiang, T. (2020). Life lacks meaning without acceptance: Ostracism triggers suicidal thoughts. *Journal of Personality and Social Psychology*, 119(6), 1423–1443. <https://doi.org/10.1037/pspi0000238>
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.
- Cuadrado, E., Tabernero, C., & Steinel, W. (2015). Motivational determinants of prosocial behavior: What do included, hopeful excluded, and hopeless excluded individuals need to behave prosocially? *Motivation and Emotion*, 39(3), 344–358. <https://doi.org/10.1007/s11031-014-9460-z>
- DeWall, C. N., Twenge, J. M., Bushman, B., Im, C., & Williams, K. D. (2010). A little acceptance goes a long way: Applying social impact theory to the rejection–aggression link. *Social Psychological and Personality Science*, 1(2), 168–174. <https://doi.org/10.1177/1948550610361387>
- Eck, J., Schoel, C., & Greifeneder, R. (2016). Coping with or buffering against the negative impact of social exclusion on basic needs: A review of strategies. In P. Riva & J. Eck (Eds.), *Social exclusion* (pp. 227–249). Springer. https://doi.org/10.1007/978-3-319-33033-4_11
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121–138. <https://doi.org/10.1037/1082-989X.12.2.121>
- Forbes. (2023, May 18). Top social media statistics and trends of 2023. <https://www.forbes.com/advisor/business/social-media-statistics>
- Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2), 161–178.
- Hayes, R. A., Carr, C. T., & Wohn, D. Y. (2016). One click, many meanings: Interpreting paralinguistic digital affordances in social media. *Journal of Broadcasting & Electronic Media*, 60(1), 171–187. <https://doi.org/10.1080/08838151.2015.1127248>
- Howell, D. C. (2005). Median absolute deviation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (1st ed.). Wiley. <https://doi.org/10.1002/0470013192.bsa384>
- Kang, R., Brown, S., & Kiesler, S. (2013). Why do people seek anonymity on the internet? Informing policy and design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2657–2666). <https://doi.org/10.1145/2470654.2481368>
- Karahanna, E., Xin Xu, S., Xu, Y., & Zhang, N. A. (2018). The needs–affordances–features perspective for the use of social media. *MIS Quarterly*, 42(3), 737–756. <https://doi.org/10.25300/MISQ/2018/11492>
- Kuehn, M. M., Chen, S., & Gordon, A. M. (2015). Having a thicker skin: Social power buffers the negative effects of social rejection. *Social Psychological and Personality Science*, 6(6), 701–709. <https://doi.org/10.1177/1948550615580170>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lee, E., Lee, J.-A., Moon, J. H., & Sung, Y. (2015). Pictures speak louder than words: Motivations for using Instagram. *Cyberpsychology, Behavior, and Social Networking*, 18(9), 552–556. <https://doi.org/10.1089/cyber.2015.0157>
- Lee, M., Kim, H., & Kim, O. (2015). Why do people retweet a tweet? Altruistic, egoistic, and reciprocity motivations for retweeting. *Psychologia*, 58(4), 189–201. <https://doi.org/10.2117/psysoc.2015.189>
- 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>
- Lutz, S., Büttner, C. M., & Neumann, D. (2023). A window to what we missed: Effects of self- versus other-exclusion on social media users' fundamental needs, emotional responses, and online coping behaviors. *Media Psychology*, 27(3), 401–427. <https://doi.org/10.1080/15213269.2023.2242769>
- Lutz, S., & Schneider, F. M. (2020). Is receiving dislikes in social media still better than being ignored? The effects of ostracism and rejection on need threat and coping responses online. *Media Psychology*, 24(6), 741–765. <https://doi.org/10.1080/15213269.2020.1799409>
- Lutz, S., Schneider, F. M., & Reich, S. (2022). Media as powerful coping tools to recover from social exclusion experiences? A systematic review on need restoration and emotion regulation through using media. *Media Psychology*, 26(4), 388–413. <https://doi.org/10.1080/15213269.2022.2147085>

- Maner, J. K., DeWall, C. N., Baumeister, R. F., & Schaller, M. (2007). Does social exclusion motivate interpersonal reconnection? Resolving the “porcupine problem.” *Journal of Personality and Social Psychology*, 92(1), 42–55. <https://doi.org/10.1037/0022-3514.92.1.42>
- Meral, E. O., van Osch, Y., Ren, D., van Dijk, E., & van Beest, I. (2021). The anticipated social cost of disclosing a rejection experience. *European Journal of Social Psychology*, 51(7), 1181–1197. <https://doi.org/10.1002/ejsp.2807>
- Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, 21(15), 3301–3307. <https://doi.org/10.1093/bioinformatics/bti499>
- Peddinti, S. T., Ross, K. W., & Cappos, J. (2014). “On the internet, nobody knows you’re a dog”: A Twitter case study of anonymity in social networks. In *Proceedings of the Second ACM Conference on Online Social Networks* (pp. 83–94). <https://doi.org/10.1145/2660460.2660467>
- Pittman, M., & Reich, B. (2016). Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. *Computers in Human Behavior*, 62, 155–167. <https://doi.org/10.1016/j.chb.2016.03.084>
- Poon, K.-T. (2018). Unpacking the mechanisms underlying the relation between ostracism and Internet addiction. *Psychiatry Research*, 270, 724–730. <https://doi.org/10.1016/j.psychres.2018.10.056>
- Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7(2), 1–14. <https://doi.org/10.1177/20563051211019004>
- Reich, S., Schneider, F. M., & Heling, L. (2018). Zero likes—Symbolic interactions and need satisfaction online. *Computers in Human Behavior*, 80, 97–102. <https://doi.org/10.1016/j.chb.2017.10.043>
- Ren, D., Wesselmann, E. D., & van Beest, I. (2020). Seeking solitude after being ostracized: A replication and beyond. *Personality and Social Psychology Bulletin*, 47(3), 426–440. <https://doi.org/10.1177/0146167220928238>
- Ren, D., Wesselmann, E. D., & Williams, K. D. (2018). Hurt people hurt people: Ostracism and aggression. *Current Opinion in Psychology*, 19, 34–38. <https://doi.org/10.1016/j.copsyc.2017.03.026>
- Rudert, S. C., Hales, A. H., Greifeneder, R., & Williams, K. D. (2017). When silence is not golden: Why acknowledgment matters even when being excluded. *Personality and Social Psychology Bulletin*, 43(5), 678–692. <https://doi.org/10.1177/0146167217695554>
- Triggs, A. H., Møller, K., & Neumayer, C. (2021). Context collapse and anonymity among queer Reddit users. *New Media & Society*, 23(1), 5–21. <https://doi.org/10.1177/1461444819890353>
- Williams, K. D. (2009). Ostracism: A temporal need-threat model. *Advances in Experimental Social Psychology*, 41, 275–314. [https://doi.org/10.1016/S0065-2601\(08\)00406-1](https://doi.org/10.1016/S0065-2601(08)00406-1)
- Williams, K. D., & Nida, S. A. (2009). Is ostracism worse than bullying? In M. J. Harris (Ed.), *Bullying, rejection, and peer victimization: A social cognitive neuroscience perspective* (pp. 279–296). Springer.
- 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>
- Yokotani, K., & Takano, M. (2021). Social contagion of cyberbullying via online perpetrator and victim networks. *Computers in Human Behavior*, 119, 106719. <https://doi.org/10.1016/j.chb.2021.106719>
- Zadro, L., Williams, K. D., & Richardson, R. (2005). Riding the “O” train: Comparing the effects of ostracism and verbal dispute on targets and sources. *Group Processes & Intergroup Relations*, 8(2), 125–143. <https://doi.org/10.1177/1368430205051062>
- Zhao, Y., Cingel, D. P., Xie, L., & Yu, M. (2023). Exploring the relations of subjective and objective Instagram use on young adults’ mental health. *Telematics and Informatics*, 77, 101921. <https://doi.org/10.1016/j.tele.2022.101921>

Curriculum Vitae

Christoph Kenntemich

Education

12/2021 – 11/2024	Research associate at the University of Kaiserslautern and Landau, Department of Social-, Environmental-, and Economic Psychology
10/2019 - 08/2021	M. Sc. Psychology, University of Giessen
10/2016 - 07/2019	B.Sc. Psychology, University of Giessen

Teaching

04/2023 – 09/2024 Empirical Project Seminar at RPTU Kaiserslautern-Landau

01/2024. – Workshop: *ChatGPT im Forschungsalltag: Nicht weniger Denken, sondern mehr.* (JuMi of the DGPs – online)

06/2024 – Workshop: *Generative KI in der Wissenschaft: Wo sind wir? Wohin kann es gehen?* (Graduate School, University Giessen, Germany)

07/2024 – Workshop: *Generative AI in the Social Sciences: Where are we now? Where can we go?* (Department of Social Psychology, University of Basel, Switzerland)

10/2024 – Workshop: *Generative AI in the Social Sciences: Where are we now? Where can we go?* (Graduate School, University of Basel, Switzerland)

Publications

Büttner, C. M., Kenntemich, C., & Williams, K. D. (2024). The power of human touch: Physical contact improves performance in basketball free throws. *Psychology of Sport and Exercise*, 72, 102610. <https://doi.org/10.1016/j.psychsport.2024.102610>

Kenntemich, C., Brückner-Collet, D. O. I., & Rudert, S.C. (2024). The Global Village's Local Rules: Cross-Cultural Differences in Online Norm-Enforcement. *Under review at Computers in Human Behavior*.

Kenntemich, C., Brückner-Collet, D. O. I., & Rudert, S.C. (2024). The Global Village's Local Rules: Cross-Cultural Differences in Online Norm-Enforcement. *Under review at Computers in Human Behavior*.

Kenntemich, C. & Rudert, S.C. (2024). The impact of conflict-discouraging codes of conduct on norm-enforcement in online forums. *Under review at Computers in Human Behavior*.

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 Acknowledgement-Seeking Behavior. *Personality & Social Psychology Bulletin*.
<https://doi.org/10.1177/01461672241297824>

Rudert, S. C., Möring, R., Kenntemich, C. & Büttner, C. M. (2023). When and why we ostracize others: Motivated ostracism in group contexts. *Journal of Personality and Social Psychology*.
<https://doi.org/10.1037/pspi0000423>

Conference Presentations

“When unsure, just ignore. How interaction norms influence the reactions towards norm-deviants.”, Conference of the European Social Cognition Network, Milan, IT, August 2022.

„...but be nice about it! How responses to norm-violators depend on interaction norms.”, General Meeting of the European Association of Social Psychology, Krakow, PL, July 2023.

“To Tweet or not to Tweet. Behavioral Responses to being rejected on Twitter.” Fachgruppen-Tagung der Fachgruppe Sozialpsychologie. Graz, AT, September 2023. (Organized Symposium)

„Too. busy to respond? The Effect of Time Constraints on Ostracism and Prosocial Decisions.” 96th Annual Meeting of the Midwestern Psychological Association (MPA), Chicago, IL, USA, April 2024

Invited Talks

“Empowered Scrolling: The Impact of Teaching Explicit Feedback Mechanisms on Social Media Experience.”, Universität des Saarlandes, November 2023.

Peer Reviews for Scientific Journals

- *European Journal of Social Psychology*
- *Group Processes and Intergroup Relations*
- *Social Influence*

Professional Memberships

- *German Psychological Society (DGPs)*
- *European Association of Social Psychology (EASP)*

Training Sessions Led

Workshop: ChatGPT im Forschungsalltag: Nicht weniger denken, sondern mehr. (JLU Giessen, Department of Social Psychology, December 2023)

Workshop: ChatGPT im Forschungsalltag: Nicht weniger denken, sondern mehr. (DGPs, Social Psychology Section, January 2024)

Workshop: Generative AI in the Social Sciences. (University of Basel, Social Psychology Section, July 2024)

Workshop: Generative AI in the Social Sciences. (University of Basel, Doctoral Program, July 2024)

Grants

- *Project: “Opfer von Deepfake: Bedrohungserleben und Dissonanz” Peer-Tandem by the German Psychological Society (DGPs) – 750€*

- Project: *“Refugee LLMs: Realizing the potential of LLMs to streamline integration.”* Funding by the German Ministry of Education and Research (BMBF) – 12.500 €

- Project: *“Refugee LLMs: Realizing the potential of LLMs to streamline integration.”* Funding by the German Ministry of Education and Research (BMBF) – 172.249,79 €

Scholarships

- Scholarship by the German Academic Exchange Service (DAAD) – Visiting Scholar

Eidesstaatliche Erklärung (III. §10 (2))

Hiermit erkläre ich, dass ich diese Dissertation eigenständig und ohne Hilfe Dritter sowie ohne Verwendung anderer als der angegebenen Hilfsmittel verfasst habe. Die verwendeten Quellen sind entsprechend gekennzeichnet. Die veröffentlichten Manuskripte und die zur Veröffentlichung eingereichten Manuskripte wurden gemeinsam mit den Co-Autoren entwickelt und wurden von keinem der Beteiligten als Qualifikationsarbeit veröffentlicht, zur Veröffentlichung eingereicht oder einem anderen Prüfungsgremium vorgelegt, auch nicht von mir an einer anderen Institution. Die Dissertation basiert auf folgenden Manuskripten:

1. Kenntemich, C., Brückner-Collet, D. O. I., & Rudert, S.C. (2024). The Global Village's Local Rules: Cross-Cultural Differences in Online Norm-Enforcement. *Under review at Computers in Human Behavior*.
2. Kenntemich, C. & Rudert, S.C. (2024). The impact of conflict-discouraging codes of conduct on norm-enforcement in online forums. *Under review at Computers in Human Behavior*.
3. 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 Acknowledgement-Seeking Behavior. *Personality & Social Psychology Bulletin*.
<https://doi.org/10.1177/01461672241297824>

Hilfsmittel: Bei der Erstellung dieser Arbeit habe ich Claude 3.5 Sonnet verwendet, um die sprachliche Flüssigkeit an wichtigen Stellen des Manuskripts zu evaluieren und zu verbessern. Nach der Verwendung dieses Tools habe ich den Inhalt überprüft und nach Bedarf überarbeitet und übernehme die volle Verantwortung für den Inhalt der Dissertation.