Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-04647-9 OPEN

Antisemitic and Islamophobic hate speech precedes a decrease in lexico-semantic diversity in comment threads online

Z. P. Rosen[1]✉ & Rick Dale[1]✉

Studies of Antisemitic and Islamophobic hate speech (AHS and IHS) demonstrate that they severely impact the psychological and social well-being of Jewish and Muslim communities. However, work to date has not adequately addressed the effect that the introduction of AHS/IHS has on subsequent expression in groups that tolerate hate speech online. We thus do not know how influential AHS and IHS are. The current study attempts to address this gap in the literature directly by providing an information-theoretic account of what happens when social media users on the website Reddit vary the intensity of Islamophobic and/or Antisemitic sentiment in their comments. We find that the more overtly Antisemitic or Islamophobic the comment, the easier it is to recover the ideas expressed in that comment from subsequent comments. In other words, comments that rank high for AHS or IHS appear to impose a strong bottleneck on the lexico-semantic diversity of subsequent conversations. This effect was strengthened after the onset of the conflict in Gaza and Israel on October 7th, 2023. Our results offer a step toward investigating how information transmission is altered due to the effects of particular kinds of HS, and have direct implications for organizations with a vested interest in content moderation.

[1] Department of Communication, University of California, Los Angeles, Los Angeles, CA, USA. ✉email: z.p.rosen@ucla.edu; rdale@ucla.edu

Introduction

Following the October 7th attacks by Hamas in Israel, and the subsequent escalation of the conflict in Gaza, reports of hate speech (HS)—specifically, Antisemitic hate speech (AHS) and Islamophobic hate speech (IHS)—sharply rose in the United States (ADL, 2023; Gamble, 2024). This increase in AHS and IHS corresponded with a stark increase in the number of hate crimes directed at Muslim and Jewish communities (ADL, 2024; Alfonseca, 2023; Allison, 2024). Much of this rhetoric was then co-opted by white supremacist groups as a recruitment mechanism for new members (Collen, 2023; Owen, 2023; Project, 2023), all but ensuring continued harassment for Jewish and Muslim communities.

As prevalent forms of HS, AHS, and IHS affect larger discourse in various ways. This includes potentially altering the observed diversity of ideas expressed in various discourses, especially when the person introducing AHS/IHS is a member of the same social group as other observers (we discuss this in the remainder of the report). The degree to which HS is capable of *altering discourse*, however, has not been studied to date. This is surprising, considering that in order to curb the proliferation of AHS and IHS online, it is necessary to first understand how they *affect* discourse as a type of speech act.

The following study sought to answer this question by directly applying a statistical framework designed to measure the flow of information between utterances made by multiple discursive participants. Specifically, we analyzed web-based discourses on the social media website Reddit. Our findings show that there is a direct relationship between how strongly a user expresses Islamophobic or Antisemitic sentiment and the ease with which the ideas couched in such comments are recoverable from subsequent comments. This "bottleneck" becomes narrower as prior comments become more easily identifiable as AHS or IHS, and emerges in the data as a decrease in lexico-semantic diversity in subsequent comments made by other users online. This effect is strongest and most consistent when looking at comments that are posted as replies to the same prior comment (i.e., *sibling* comments). Furthermore, the presence of either AHS or IHS in comments written after October 7th, 2023, predicted even stronger convergence in sibling comments and direct replies, suggesting that events in Israel and Gaza may have increased the influence of AHS and IHS in online discourse.

Hate Speech as a speech act. In order to understand how HS can affect subsequent discourse, we need to understand HS as a kind of speech act. In his initial definition of speech acts, J.L. Austin (1975) wrote that speech acts consist of three core components: their semantics (*locutionary force*), the action that a speaker wants to accomplish by saying the utterance (*illocutionary force*, or the speaker's *illocutionary intent*), and the actual effect an utterance has on the social environment (*perlocutionary force*). It is worth noting that speech acts can act as unique *referential propositions* as well (Searle grouped such expressions under the moniker of *expressive* speech acts in his hierarchy based on the ways they "expressed" interlocutors' presuppositions. See: Searle, 1975). In such cases, a speech act serves as a proposal for how speakers should refer to some object or idea, thus proposing specific means of conceptualizing the relationship of those things—a kind of "schema"—to the interlocutors being addressed (Clark and Wilkes-Gibbs, 1986; Enfield and Sidnell, 2022).

The effects of HS on target populations—one of its perlocutionary effects—is well described in the literature. Individuals targeted by HS report long-term anxiety and depression (Saha et al., 2019; Tynes et al. 2008), as well as being at an increased risk of suicide ideation and suicidality (Marshal et al., 2011). HS also has a chilling effect on communication and innovation in organizational cultures (Leets and Giles, 1999), can cause victims to avoid public engagement after exposure to it (Henson et al., 2013), and decrease the public's empathy toward its targets (Pluta et al., 2023).

This focus on the perlocutionary effects of HS on targeted populations has guided the implementation of several automated systems to date for the detection of HS online. Indeed, some machine learning classifiers focus on HS that target minority groups specifically. A prominent classifier with this focus was chosen to assist our analysis by automatically producing a Hate Speech rating (HS rating) of IHS and AHS (Vidgen et al. 2021). Expanding on past work, Vidgen et al. (2021) use the following criteria to identify examples of HS directed toward minority groups: (1) speech that is explicitly derogatory based on minority status, (2) speech that expresses animosity toward minorities, (3) speech that threatens individuals or groups based on their minority status, (4) speech that expresses outright support for hate proliferating entities, and (5) speech that dehumanizes individuals from a minoritized community. Vidgen et al.'s definition is thematically consistent with views expressed by other scholars who study HS in terms of speech act theory and intergroup communication (Calvert, 1997; Rae, 2012).

The existing literature has not addressed HS as an expressive (or *referential*) speech act to date. This is somewhat surprising. We know that other forms of potentially harmful speech can fundamentally alter the way individuals behave and spread information within social networks (Beknazar-Yuzbashev et al., 2022). Various groups online will not only differentially engage in HS, but in different kinds of HS (Rieger et al., 2021), and some recent work proposes that feedback loops (which may include authors observing the adoption of their hateful ideas in subsequent comments) in online engagement can drive an increase in the degree of hatefulness expressed by individuals (Walther, 2022). In addition to this, repeated exposure to HS in particular can cause individuals to express themselves more hatefully over time (Bäck et al., 2018), which further supports the idea that HS is used as a referential speech act by showing that it becomes a more dominant mode of expression as an individual increasingly accepts it as a legitimate means of expression. From the perspective of speech act theory and pragmatics, we really should put more focus on unpacking how HS alters referential understanding within groups.

Whether or not individuals propagate referential content taken from hateful comments may yield important insights into "information flows" in social media. Ozalp et al. (2020) found strong evidence that on Twitter (now X) tweets containing counter-hate messages, authored by trusted organizations, had longer duration (were retweeted across longer time periods) and spread (were retweeted by more individuals) than biased, hateful messages on Twitter as measured in the number of retweets. While their results show that social media users *endorse* HS at a given rate on X.com (even if that rate is lower than counter-HS), it does not directly measure how specific ideas might spread rhetorically within a group. This is critical, however, for understanding the full reach of HS online. Groups of individuals may not always repost, quote, or retweet a message. They will, however, respond to hate-espousing content and modify the language they use based on their assessment of it and the social context surrounding it (Soliz et al., 2021)—positive or negative.

When HS serves as a referential speech act it should have at least one observable perlocutionary effect in intragroup discourse —others should adopt or extend the specific concepts/ideas espoused within it. In other words, the perlocutionary effect of HS as a referential speech act should be that it "bottlenecks" the

diversity of expression observed in subsequent speech in ways that are predictable based on the content, or schema, proposed in the hateful comment.

The adoption of a specific schema is a question of what is called "convergence"—a concept that we borrow from Communication Accommodation Theory (CAT) and will discuss in section "CAT, Convergence, and HS".

Defining Antisemitic and Islamophobic hate speech. AHS and IHS are subtypes of HS that are specifically directed toward Jewish (AHS) and Muslim (IHS) communities and individuals. Over AHS and IHS, we define whether an utterance meets both of the following two conditions: (1) the utterance fits the definition of HS as expressed in the previous section, and (2) the utterance *explicitly* foregrounds Jewish or Muslim people by using language that references one of these groups.

For the purpose of measurement, these two conditions can occur independently. For example, a comment might express HS, but not foreground Jewish people and would thus not be an example of AHS. A comment might also foreground Jewish people and not express a hateful sentiment, which would also preclude it from being an example of AHS. Consider, the following examples: [1]

(1) **Foregrounds Jewish communities/individuals**

a. "The **Khazarians** merely identify themselves as Jews." (AHS) *Uses the dog-whistle "Khazars" to refer to European Jewish communities and describes them as being "false Jews."*

b. "The French Antarctic Expedition (circa 1908–1910) first spotted the island from afar, and Charcot named it **Rothschild** Island to honor Baron Edouard de **Rothschild** (1868–1949)."
(not AHS) *Describes why an island was given the name "Rothschild" without necessarily instantiating antisemitic conspiracy theories about the Rothschild family*.

(2) **Foregrounds Muslim communities/individuals**

a. "One of the worst scenarios is being **both Muslim and uneducated**."
(IHS) *Equates being Muslim with being inferior*.

b. "For those curious, this sound is **the Adhan, or call to prayer**, broadcast five times daily in all **Muslim-majority** countries."
(not IHS) *Describes the call to prayer in Muslim religious traditions*.

(3) **HS (not AHS or IHS)**
"What's up you **autists**??"
(HS) *Uses the word "autists" as a derogatory term for individuals with mental disabilities*.

Note that in the HS espousing lines for comments foregrounding Jewish and Muslim communities, the ideas being expressed are derogatory in nature. In example 1.a (the AHS espousing example), the text uses a dog whistle for European Jewish communities associated with claims that European Jews are not actually Jewish, as well as stating the antisemitic meaning of the word "Khazar" directly. In example 2.a (the IHS espousing example), the author states that being Muslim is a "bad scenario" based on religious identity. In both examples, the comments meet criteria (1) from Vidgen et al. (2021)' definition of HS—both examples are explicitly derogatory based on the minority identity of the groups being foregrounded. Additionally, both comments not only meet that criteria but also explicitly foreground Jewish and Muslim communities, meeting our specific definition of AHS and IHS. There is a little similarity between the non-HS examples for both, besides that they state a seemingly simple fact while still foregrounding Jewish and Muslim communities (though there may still be Islamophobic or Antisemitic intent implied by their context). And while example 1.a, example 2.a, and example 3 are all explicitly derogatory in their discussion of various minority identities, example 3 foregrounds derogatory views of intellectual abilities and neurodivergence and not Jewish or Muslim communities, thus rendering it HS but not AHS or IHS.

Based on this definition, then, it makes sense to treat AHS and IHS as *additive* in a quantitative framework. This would mean that one would measure IHS and AHS via the combined effects on some discourse of (1) the degree to which an utterance expresses HS, (2) whether that same utterance foregrounds either Jewish or Muslim people, and (3) any interaction of these two factors.

CAT, convergence, and HS. First proposed in the 1980s, CAT (CAT: Soliz et al., 2021) describes the interaction of linguistic patterns and social behavior. It operationalizes convergence—the tendency for speakers to adopt similar means of expression—and its opposing process, divergence, in terms of several linguistic features. These include phonetic/phonological convergence (Lewandowski and Jilka, 2019; Manson et al., 2013), syntactic convergence (Muir et al., 2016), and convergence on specific lexical expressions as the basis of group identity (Bradac et al., 1988; Hilte, 2023), which in most cases yields, as a result, alignment in individuals' semantic representations of key ideas and events (Brennan and Clark, 1996; Pickering and Garrod, 2004). Convergence typically indexes social proximity among speakers, positive affect, and co-group membership (Soliz et al., 2021). Lexico-semantic convergence has also been shown to be important for coordination between interlocutors (Alviar et al., 2023; Coco et al., 2018).

If they exist, "bottlenecks" in lexico-semantic diversity caused by AHS and IHS are probably driven by convergence. In communities where HS is more tolerated, convergence may occur more deeply when messages contain hateful elements. Thus, when comments containing AHS and IHS are introduced to discursive exchanges we predict that the lexico-semantic content of messages that include them should be converged to more deeply than the lexico-semantic content of messages that do not.

Quantitatively measuring convergence. There are a handful of quantitative methods that can be used to measure convergence in dialog (Srivastava et al., 2025). A particularly powerful quantitative measurement of convergence is convergence-entropy (CE: Rosen and Dale, 2023). At a high level, CE replicates a kind of Shannon experiment. As a speaker generates an utterance, a listener asks whether the lexico-semantic content of the speaker's utterance could be predicted by having been exposed to the lexico-semantic content of an utterance made by either another individual speaker or by the members of some social group. Convergence is thus defined as when more of the speaker's utterance could be predicted from prior exposure to other utterances and can be measured in terms of entropy (where convergence is equated to low entropy, or low CE). The authors then show that it is possible to estimate CE using a kind of language model that represents lexical meaning spatially in a high-dimensional vector. They specifically use a transformer language model to do so (Brown et al., 2020; Devlin et al., 2019), which as a class of models has been shown to correlate well with human semantic processing (Goldstein et al., 2022; Nishida et al., 2021). Because of its generalizability and lack of a priori assumptions, we use CE as a measurement of convergence in the current study.

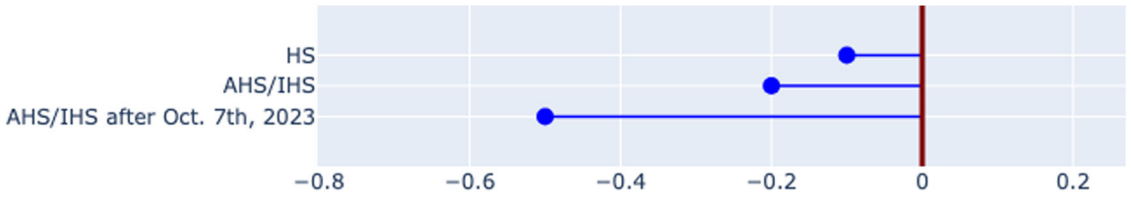


Fig. 1 Visualization of the hypothesized direction and magnitude for effects of HS on CE.

When some utterance *x* precedes an utterance *y*, CE measures how much of the older utterance *x* can be replicated after reading the latter utterance. Thus, it is interpreted as measuring how much the latter utterance *y* converges with the content expressed in the prior utterance *x*. The lower the CE, the more information about *x* can be recovered from *y*, and thus the higher the convergence of the utterance *y* to the ideas expressed in the utterance *x*.

CE, as described here, can be thought of as a measurement of the degrees of freedom that the author of comment *y* would have to exercise to yield the ideas they expressed when compared to ideas originating in comment *x*. This intuition is consistent with Claude Shannon's interpretation of entropy, who wrote that "[quantities of entropy] play a central role in information theory as measures of information, choice and uncertainty" (Shannon and Weaver, 1949). For Shannon, the idea that entropy was a measurement of how much "choice" was exercised by individuals encoding some message was a central consideration of his definition, and it also features prominently in our conceptualization of convergence in the current project.

This leads us to our first hypothesis:

> ***Hypothesis 1:*** We predict that CE will be lower when a prior comment that expresses overt IHS or AHS (i.e., a comment expresses HS and foregrounds either Jewish or Muslim folks explicitly) is compared to later comments.

Prior research has also shown that convergence tends to increase following sensational news (Hiaeshutter-Rice and Hawkins, 2022). And "galvanizing events" have in the past triggered waves of Antisemitic hate (Ozalp et al. 2020). Based on the previously mentioned reporting surrounding AHS and IHS at the moment, we have reason to believe that the conflict in Gaza may affect how deeply statements containing IHS and/or AHS are converged to. We thus also predict the following:

> ***Hypothesis 2:*** We predict that CE will be lower when the prior comment is (1) high in either AHS or IHS and (2) is written after October 7th, 2023.

A chart demonstrating our hypotheses as listed here is included in Fig. 1 as a visual reference.

To better understand what our hypotheses look like in actual discourse, consider the following exchange between four interlocutors in an online message board. Example 4 demonstrates an Islamophobic comment *x* in the first line, and three comments made to the same thread after it: a subsequent comment that rebukes the author of the comment *x*, a subsequent comment that "converges" with the ideas expressed in the comment *x*, and a subsequent comment that expresses IHS as well but *does not converge* with the comment *x*. Each example is labeled with the calculated CE for the comparison of the initial comment *x* to the subsequent comments (i.e., comparing the topmost sentence *x* to example 4.a, comparing *x* to example 4.b, etc.).

(4) "Every time I pass by the local Arab sales booth, *their* music fucking bothers me. It is just a loud, shrill flute sound without any melody."

IHS: *Describes the music choices of a specific sales booth using language that is derogatory toward Muslim communities.*

a. "With all due respect, what the shit are you talking about?"
Not HS: *Openly rebukes the author of the comment x*
(*CE* = 6.91)

b. "Why do Arabs prefer music that lacks melody and rhythm? They all just sound like strange noises."
IHS: *Extends the prior comment x to all ethnic Arabs and expands on the derogatory language used to describe Muslim-coded music.*
(*CE* = 4.52)

c. "It is necessary to prohibit Islam."
IHS: *Does not engage with the ideas expressed in the comment x and instead calls to ban Islam.*
(*CE* = 6.57)

While example 4 and example 4.c both express Islamophobic sentiments, they do so by invoking very different ideological framings of Islam. The first sentence in example 4—the comment *x*—is a derogatory comment on the quality of music perceived as being Muslim, while example 4.c is an outright cry to ban Islam as a religion. Both are IHS, but they do not *converge* in how they express Islamophobic sentiment. Meanwhile, the comment *x* (example 4) and example 4.b both use derogatory language to describe music perceived as being Muslim, even though they are not expressing identical ideas—comment *x* focuses on degrading the music played in a single shop, while example 4.b extends the same derogatory sentiment to all ethnic Arabs and claims that Muslim-coded music lacks an additional feature of rhythm. While they are not the same, example 4.b is building on the same set of core ideas as those expressed in the comment *x*, making it easier to recover the ideas expressed in *x* from example 4.b. As a result, CE is the lowest between these two examples. As a general measurement tool, CE is sensitive not just to *linguistic mimicry*, but to lexico-semantic, or *conceptual*, similarities expressed in different utterances.

Reddit, interaction types, and HS in online discourse. While several social media sites may act as a medium for disseminating HS, Reddit is unique in that groups (known as *subreddits*) are self-selecting and somewhat insular. Every subreddit has its own distinct membership, conventions, and demographic composition (Shatz, 2017). These factors mean that Reddit provides a useful window into *intragroup* communication online. Thus, it is an excellent laboratory for studying not just how AHS and IHS affect discourse on the web, but how they affect the expressions used by other members of groups that are allegedly more tolerant of them.

On Reddit, users (called *redditors*) are afforded two options with regard to how they can interact with the content posted by other users. A redditor can (1) directly reply to a comment (in which case, the comment a user replies to is the *parent* to their reply) or (2) as part of a longer thread of comments replying to the same parent comment (i.e., the older comment and the new

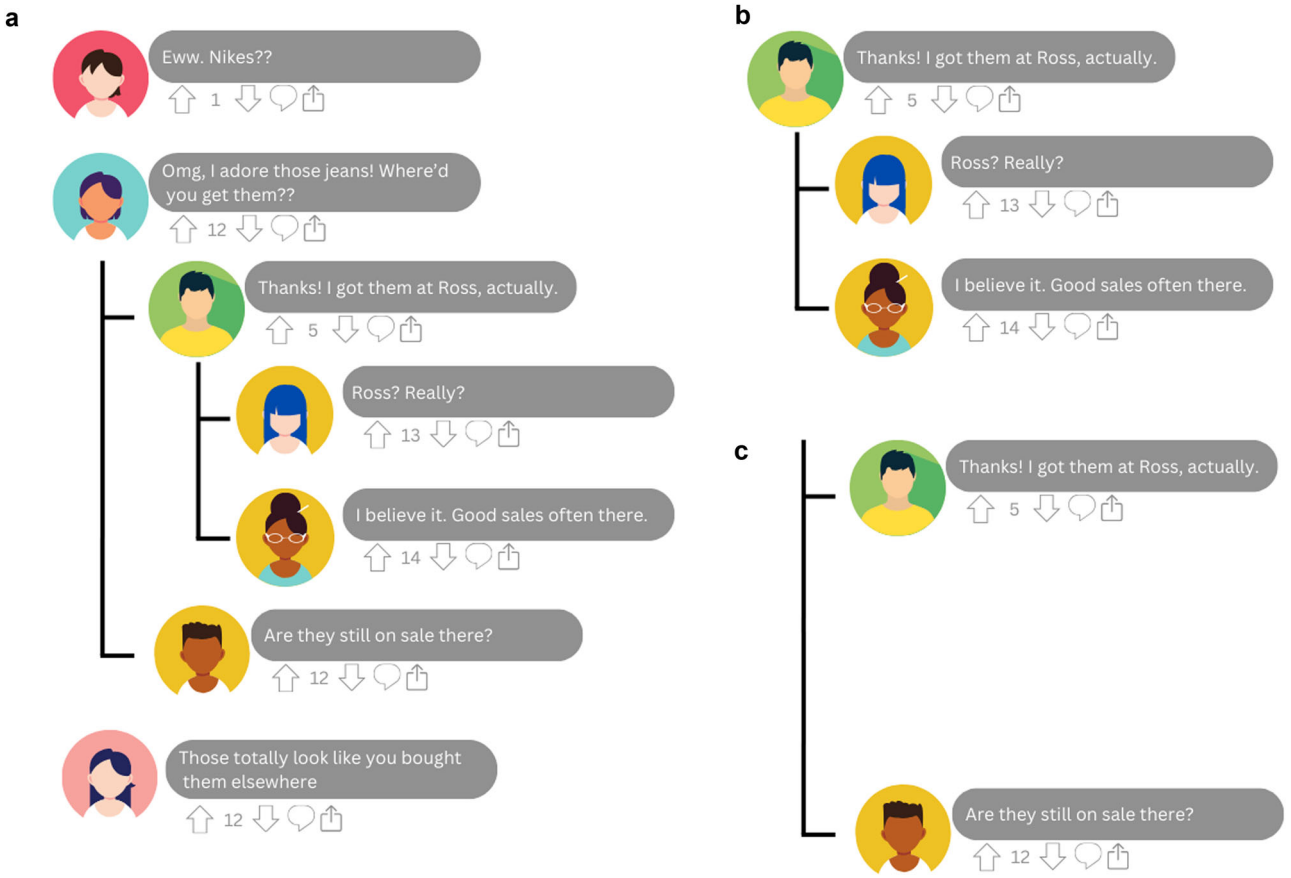


Fig. 2 Visualization of possible relationships between comments in a Reddit submission. **a** Structure of a full submission on Reddit. **b** A "parent" comment and two replies. **c** Two "sibling" comments.

comment are *siblings*). A visualization of these relationships is given in Fig. 2 for reference.

The study of the dynamics of individuals posting HS on social media is still in its infancy. Even so, studies show that HS tends to cluster in specific loci in online spaces (Cinelli et al., 2021; Inwood and Zappavigna, 2023). One important observation is that "haters" will often post HS as a non-exclusive group activity referred to as "hate parties" (Rea et al., 2024). For hate parties, the illocutionary intent of inflicting psychological harm on the targets of HS is secondary to posting similar messages *en masse* to signal epistemic solidarity. Rea et al. invite readers to think of the dynamics of a hate party as individuals "hanging out" together in the same comment threads online, like people "clustered together in conversation in different parts of a room at the same party." This difference in intention means that similarly hateful messages cluster together as responses to a particular parent's comment.

While hate parties *can be maintained* by several members of a single group interested in continuing a particular thread in public (Rea et al., 2024), hate parties are not defined by their composition—a "hate party" does not need to be composed of highly connected members of the same group. There is no requirement for individuals to have *any* history with one another for them to gain admission to a hate party. The only "entrance requirement" is that individuals engaging in a hate party need to be interested in affirming a particular, hate-based view with other partygoers. In the words of Rea et al.:

> "[Hate parties] could be more accurately characterized as broader discussions taking place simultaneously and interactively in different online social spaces, like groups of people clustered together in conversation in different parts of a room at the same party. The lack of any clear direction or coordination can be explained in large part by the differences between hate raids and hate parties regarding their respective goals. While the purpose of a hate raid is to disrupt and harass a specific target or targets, we want to suggest that hate parties are not motivated by antagonism; their goal is to share and promote hateful world views …" (Rea et al., 2024).

The idea of a "hate party" is, thus, not that there is a "party" of individuals continuously engaging with one another, who selectively coordinate among themselves when and where they will congregate. If hate parties are parties, they are not exclusive. A hate party is a designation—by some signal—of a locus embedded in an ongoing discourse wherein people can congregate and engage with the kind of hate-based content of their choosing, regardless of personal history.

The image of "people clustered together in conversation in different parts of a room at the same party" is particularly apt, even in intragroup settings like those explored in the current study. While it is not explored in detail in Rea et al. (2024), in intragroup discourse where many different ways of expressing hateful sentiment could occur within the same submission, one can imagine individuals clustered together not just according to whether prior comments expressed hateful or congenial-to-hate sentiment, but according to the *kind* of sentiment that they were expressing. If one is interested in making derogatory comments about the musical preferences of different minority groups (as in example 4) one can pick a "room" in a Reddit post where other partiers are engaged in similar discussions. This is in fact at the

heart of Hypothesis 1, and is exemplified in the difference between the comment *x* and example 4.b in example 4.

It is unclear what a signal to congregate might look like for a hate party. Is it that, like a mother duck protecting her ducklings, a member of some group posts an initial hateful comment and other individuals, interested in engaging with the hateful ideas of that author, post *replies* under the umbrella of that initial comment (i.e., the *parent* comment is hateful and that signals others to join the "hate party?"). If that is the case, one would expect that parent comments expressing AHS/IHS are strongly converged to in discourse. Or is it that once one individual posts a hateful reply, other individuals will post related and/or supportive comments in the same thread of replies (i.e., a hateful, older *sibling* comment signals others to join the "hate party")? If this is true, we would expect strong convergence of more recently posted sibling comments to older sibling comments that express AHS/IHS, but we may or may not observe strong convergence between hate-espousing parent comments and replies to them—strong convergence between parent comments and replies is not entailed by strong convergence between newer and older sibling comments. Maybe both are at play? Effects arising from *engagement strategy*—like content related to hateful ideas clustering in specific *threads* of sibling comments—should be taken into consideration when assessing the effects of HS on convergence. Otherwise, we risk missing an important difference in how information flows are affected by preceding Antisemitic or Islamophobic sentiment.

We weighed this heavily in our final analysis. While it did not change our initial hypotheses, it led us to test our hypotheses in parent comments and sibling comments separately.

Materials and methods

Materials: data and data collection. All data were collected using the PRAW Reddit API package in Python. Code to replicate data collection and "rehydrate" our comments can be found at https://bit.ly/AHS_IHS, and can be used to replicate our results.

Data was collected in two phases. First, we identified several candidate subreddits to search for examples of AHS and IHS by scraping mentions of specific groups in the subreddit r/AgainstHateSubreddits (see Hickey et al., 2024, for other studies that have leveraged this subreddit in the past as well). We selected the ten most mentioned subreddits for categories of Antisemitism and Islamophobia. We then crafted a Lucene query for AHS and IHS (separately) using the lexical resources provided by the Weaponized Word Project[2] and searched our list of subreddits. We added to our IHS Lucene query the words "Muslim", "Islam", and "Palestine" to cover additional mentions of Muslim people and communities. We also added the terms "Jew", "Judaism", and "Israel" to our AHS Lucene query for the same reason.

Two final preprocessing steps were performed after data collection. To identify whether or not a comment explicitly foregrounded Jewish or Muslim communities, we converted the key terms from our Lucene query to separate regular expressions for IHS and an AHS and applied them to each comment. We then ran each comment through an HS classifier to yield a linear rating for how hateful a comment is (see "HS classifier and HS ratings").

A diagram depicting the process of data collection is provided in Fig. 3 for reference. Our exact queries and lexical search terms can be found online at https://bit.ly/AHS_IHS.

An important note needs to be made at this junction: the subreddits we studied have a history of being *accused* of generating AHS or IHS. That does not mean that members of these groups are universally or uniformly hateful. Nor does it mean that those accusations are always founded, or that these subreddits are the only sources of IHS or AHS on Reddit. We discourage any such reading of our data collection process. Rather, due to the sheer size of the Reddit ecosystem, mentions of potentially hateful submissions and comments posted in r/AgainstHateSubreddits constitute a useful starting point when identifying places to search for examples of such behavior.

Quantitative model description

Formalizing convergence measurement. Calculating CE, using word vectors, is accomplished in a few steps. First, all tokens *i* in an utterance *x* are converted to word vectors (E_{xi}) using a transformer language model. The process is repeated for all the tokens *j* in a second utterance *y* (yielding E_{yj}). To answer the question of how much one can recover of the meaning of the utterance *x* just from using the content in the utterance *y*, the researcher then compares the word vector for each token *i* in the comment *x* to all the tokens *j* in the comment *y* and selects the token *j* whose word vector has the lower Cosine Error (CoE) to the word vector for the token *i*. CoE is a measure of *distance*, but it can be used to calculate a probability $P(E_{xi}|E_y)$ representing how likely it is that two-word vectors "mean" the same thing—if the CoE for two-word vectors is 0, that would indicate that the two-word vectors are the same, and thus, is interpreted in computational linguistic studies as the tokens sharing the same lexico-semantic meaning. This conversion to a probability is accomplished by applying a Half-Gaussian distribution with location $\mu = 0$ and a free parameter for the scale σ, which codifies the idea that if CoE is 0 (i.e., there is no CoE), then this would be strong evidence that the two words, as encoded in word vectors, mean the same thing.

$$P(E_{xi}|E_y) = P_{\mathcal{N}_{[0,\infty]}}\left(\min_j(CoE(E_{xi}, E_{yj}))\middle|\mu = 0, \sigma\right) \quad (1)$$

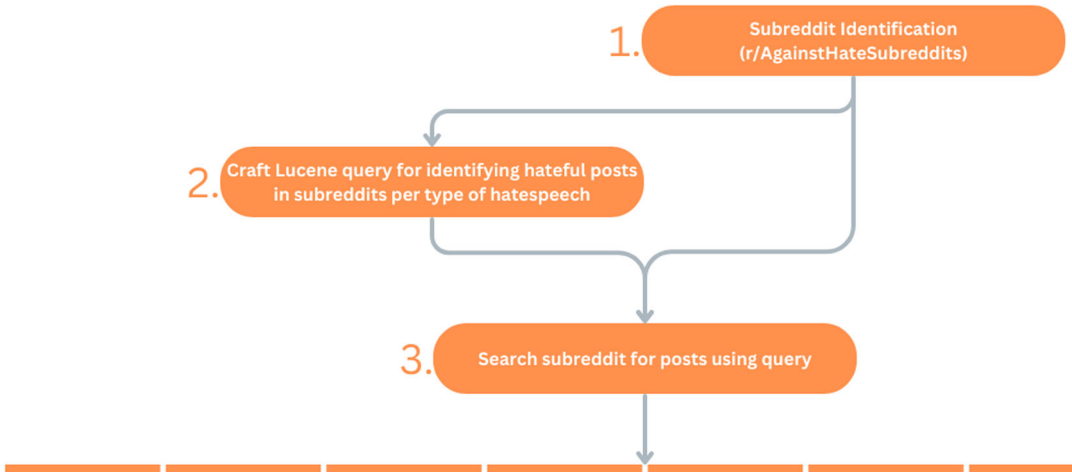
The probability rendered for each token is then used to estimate the total entropy for the sequence of tokens, as shown in Eq. (2).

$$H(x; y) = -\sum_i P(E_{xi}|E_y)\log P(E_{xi}|E_y) \quad (2)$$

Word vector model. We used the "roberta-base-uncased" model (Liu et al. 2019) to generate word vectors for the current study. The word vector model is freely available on HuggingFace's library of transformer models (Wolf et al. 2020). This model was selected because it has been shown to perform well on several semantics-related tasks in NLP, while not being specialized to a particular domain. No fine-tuning of the model was performed before the study.

Model parameters. CE has a single free parameter—σ—which defines the scale or "slope" of the half-Gaussian distribution used to convert CoE values to probabilities. We opted to model a "naive" listener who would accept a wide range of similarity values as indicative of true similarity in meaning, setting σ to be equal to 1.5.[3] This naive value of σ effectively biases our results towards a null hypothesis, as it gives significant probability to a wide range of possible CoE values.

HS classifier and HS ratings. For the amount of data we collected we opted to generate linear ratings for how hateful comments were using an automatic system for labeling HS. We specifically leveraged the HS classifier created by the Facebook (now Meta) AI Research team, as detailed in Vidgen et al. (2021), as it achieved high performance on classifying examples of HS from a number of HS datasets in English ($F1 = 92.01 \pm 0.6$), as well as performing well during active adversarial testing by human annotators (annotators were tasked with writing examples for



comment_id	user (ID)	body	comment_ups	comment_parent	comment_created_at	...	submission_id
1	101	"I believe..."	1	None	1697318293	...	Xb-10
2	115	"I disagree!"	4	1	1697318572	...	Xb-10
...	...	...	...	...	...	...	...
N	312	"Let's table this..."	-1	4	1697319000	...	Xc-111

Fig. 3 Visualization of the three steps involved in collecting data for the current study. While the columns shown here indicate values taken from Reddit comments directly, all corpus resources are stored as indices for specific submissions, consistent with Reddit's data collection and sharing policy.

hate and non-HS to attempt to get the model to misclassify their utterances).

Results

We analyzed IHS and AHS separately. There were a couple of reasons for this decision. First, only a single group overlapped between both conditions—r/4chan. Second, there were significant differences in rates of discussions of either target group based on whether data was found while searching for examples of IHS or AHS ($\chi^2(1) = 1270.36$, $p < 1e^{-9}$). In other words, discussions of Muslim people and communities did not tend to occur in the same submission/posts as discussions about Jewish people and communities.

Details about each subreddit studied for examples of IHS and AHS, including the total number of submissions whose comments were analyzed, the total number of comments analyzed per each group, and the total number of pairwise comparisons made when calculating CE values for comments pulled from each subreddit are reported in Table 1.

One subreddit identified for potential AHS examples and three subreddits identified for potential IHS examples (4 subreddits total) had been either banned or quarantined on Reddit at the time of our study and were thus excluded. Four out of the seven remaining subreddits identified for IHS contained a large number of non-English utterances (i.e., r/indiaspeaks, r/dankinindia, r/indiandankmemes, and r/hindutvarises). Due to the specific word vector model we used for our analyses (the 'roberta-base-uncased' was not trained on multi-lingual utterances), we excluded these subreddits from our final analysis. This left 3 subreddits in our sample for IHS and 9 subreddits for AHS.

IHS. For direct replies to a comment x (i.e., the comment x is the parent of the comment y), the HS rating of comment x was a significant predictor of lower CE (-0.394; $z = -4.82$, $p < 1e^{-5}$, 95% CI: [-0.554, -0.234]). However, whether the comment x foregrounded Muslim communities was not a significant predictor of CE (-0.0519; $z = -0.239$, $p = 0.811$, 95% CI: [-0.478, 0.374]). The interaction of these two variables was not a significant predictor of CE (0.194, $z = 0.51$, $p = 0.610$, 95% CI: [-0.552,0.971]). The combined effect of the HS rating of the comment x, the comment x foregrounding Muslim communities, and the interaction of those two variables (IHS) was not a significant predictor of lower CE values (Wald Test: $F(1; 11, 155) = 0.907$, $p = 0.341$). IHS expressed after October 7th was the strongest predictor of lower CE (Wald Test: $F(1; 11, 155) = 5.64$, $p = 0.0175$).

For siblings to a comment x, both the HS rating of the comment x and whether the comment x foregrounded Muslim communities were significant predictors of lower CE (-0.523; $z = -8.15$, $p < 1e^{-9}$, and 95% CI: [-0.561, -0.308] and -1.21; $z = -5.76$, $p < 1e^{-5}$, and 95% CI: [-1.61, -0.792], respectively). The interaction of these two variables was not a significant predictor of CE (0.46; $z = 1.38$, $p = 0.168$, and 95% CI: [-0.216, 1.09]). The combined effect HS rating of the comment x, foregrounding Muslim communities, and the interaction of those two (IHS) was a significant predictor of lower CE values (Wald Test: $F(1; 228, 131) = 29.88$, $p < 1e^{-5}$). IHS expressed after October 7th was the strongest predictor of lower CE (Wald Test: $F(1; 228, 131) = 5.75$, $p = 0.0165$).

Parameter estimates for individual parameters and their significance are given in Table 2. Parameter estimates, Z-statistic values, and full p-values for model estimates are available at https://bit.ly/AHS_IHS. A visualization of the magnitude of the effects of HS, IHS, and IHS expressed after October 7th for both direct replies and sibling comments is provided in Fig. 4.

AHS. For direct replies to a comment x (i.e., comment x is the parent of comment y) the HS rating of the comment x and whether it foregrounded Jewish people and communities were

Table 1 Data characteristics for all subreddits analyzed.

Subreddit		Submissions	Total comments
IHS			
Fingmemes		73	3106
2westerneurope4u		100	7891
Indiaspeaks		38	2310
Dankinindia		92	4443
4chan		97	8077
Indiandankmemes		97	4524
Hindutvarises		30	328
AHS			
Israelexposed		71	1194
Tucker_carlson		68	766
Greentext		97	8569
Timpool		75	2182
Conspiracy		95	18,981
Politicalcompass		88	3267
Conspiracy_commons		97	6207
4chan		98	5762
Wallstreetsilver		97	9365
Corpus characteristics by relationship between comments			
		# *x*	# *y*	*n* combinations
(IHS)	*Preceding comment (x) is the parent of the subsequent comment (y)*	7807	12,714	12,714
	Preceding comment (x) is a sibling of the subsequent comment (y)	10,666	10,664	240,623
(AHS)	*Preceding comment (x) is the parent of the subsequent comment (y)*	25,624	40,655	40,655
	Preceding comment (x) is a sibling of the subsequent comment (y)	28,953	28,885	587,567

Table 2 Parameter estimates for IHS LME analysis.

	x is the parent of *y*	*x* is a sibling of *y*
Intercept	−0.87****	−1.65****
x likes received	$2.62e^{-4}$	$2.25e^{-5}$
After October 7	−0.0342	−0.138*
x HS rating	−0.395***	−0.434****
x about Muslim people	−0.0519	−1.2****
x HS rating:x about Muslim people	0.194	0.435
After October 7: x HS rating	−0.06	0.132
After October 7: x about Muslim people	−0.0189	0.633
After October 7: x HS rating: x about Muslim people	−1.47	−0.736
Comment Δ	−0.0138**	$1.8e^{-4}$**
nx	0.192****	0.211****
ny	−0.0163****	−0.00365****
1 | subreddit	0.00526	0.207****
1 | x user	$-2.97e^{-6}$	$8.9e^{-7}$
1 | y user	$-4.06e^{-7}$	$1.64e^{-6}$**
1 | x	0.116****	1.72****

* = p <0.05, ** = p <0.01, *** = $p < 1e^{-5}$, **** = $p < 1e^{-9}$.

Table 3 Parameter estimates for AHS LME analysis.

	x is the parent of *y*	*x* is a sibling of *y*
Intercept	−1.15****	−1.36****
x likes received	$8.51e^{-5}$	$-6.27e^{-7}$
After October 7	−0.25**	−0.0113
x HS rating	−0.555****	−0.298****
x about Jewish people	−0.62****	−0.734****
x HS rating:x about Jewish people	1.16****	0.516***
After October 7:x HS rating	−0.166	−0.427**
After October 7:x about Jewish people	0.0715	−0.0404
After October 7:x HS rating:x about Jewish people	−0.642	0.335
Comment Δ	0.00883***	$2.76e^{-5}$
nx	0.187****	0.192****
ny	−0.0172****	−0.00434****
1 | subreddit	0.0142	0.0166***
1 | x user	$1.62e^{-6}$	$7.79e^{-7}$
1 | y user	$-2.09e^{-6}$	$8.9e^{-7}$***
1 | x	0.13****	2.81****

** = p <0.01, *** = $p < 1e^{-5}$, **** = $p < 1e^{-9}$.

both significant predictors of lower CE (−0.555; $z = -7.11$, $p < 1e^{-9}$, 95% CI: [−0.708, −0.402] and −0.620; $z = -7.34$, $p < 1e^{-9}$, and 95% CI: [−0.786, −0.454], respectively). The interaction of these two variables was a significant predictor of increased CE (1.16, $z = 6.15$, $p < 1e^{-9}$, 95% CI: [0.791, 1.53]). The combined effect of the HS rating of the comment *x*, it foregrounds Jewish people and communities, and the interaction of those two variables (AHS) was not a significant predictor of CE values ($F(1;\ 35,\ 783) = 0.0097$, $p = 0.921$). However, AHS expressed after October 7th was the strongest predictor of lower CE (Wald Test: $F(1;\ 35,\ 783) = 10.46$, $p = 0.00122$).

For siblings to a comment *x*, the HS rating of the comment *x* and whether it foregrounded Jewish people and communities were both significant predictors of lower CE (−0.298; $z = -6.93$, $p < 1e^{-9}$, 95% CI: [−0.382, −0.214] and −0.734; $z = -14.50$, $p < 1e^{-9}$, and 95% CI: [−0.833, −0.635], respectively). The interaction of these two variables was a significant predictor of an increase in CE (0.516; $z = 4.72$, $p < 1e^{-5}$, 95% CI: [0.302, 0.731]). The combined effect of the HS rating of the comment *x*, foregrounding Jewish people and communities, and the interaction of those two variables (AHS) was a significant predictor of lower CE values ($F(1;\ 557,\ 200) = 42.25$, $p < 1e^{-9}$). AHS

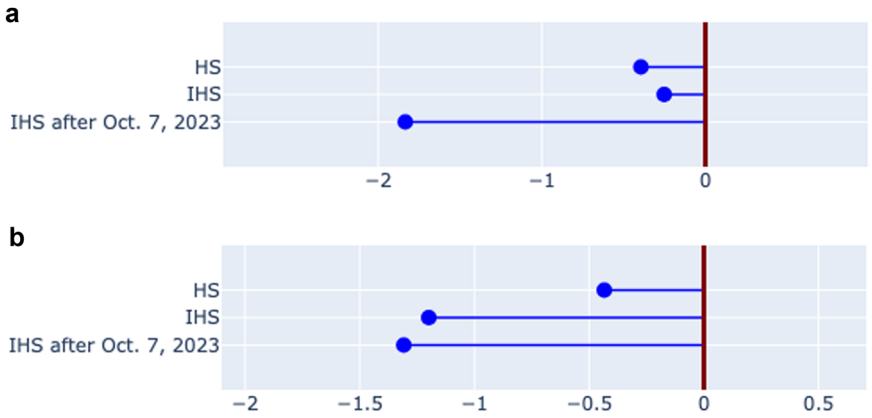


Fig. 4 Visualization of the effects of HS, IHS, and IHS after October 7th for parent comments compared to their direct replies, followed by sibling comments. **a** Parameter estimates for parent comments and their direct replies. **b** Parameter estimates for sibling comments.

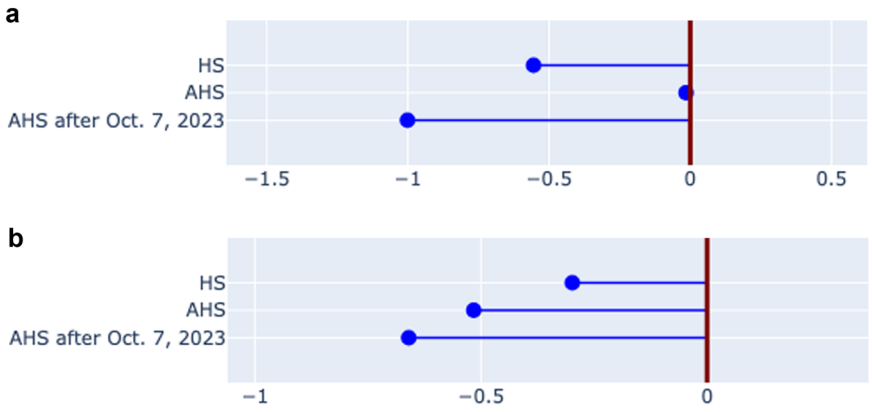


Fig. 5 Visualization of the effects of HS, AHS, and AHS after October 7th for parent comments compared to their direct replies, followed by sibling comments. **a** Parameter estimates for parent comments and their direct replies. **b** Parameter estimates for sibling comments.

expressed after October 7th was the strongest predictor of lower CE ($F(1; 557, 200) = 15.84$, $p = 6.88e^{-5}$).

Parameter estimates for individual parameters and their significance are given in Table 3. Parameter estimates, Z-statistic values, and full p-values for model estimates are available at https://bit.ly/AHS_IHS. A visualization of the magnitude of the effects of HS, AHS, and AHS expressed after October 7th for both direct replies and sibling comments is provided in Fig. 5.

Discussion

Our findings show that effects on convergence (and thus on "information flows") arising from the degree to which comments express overt IHS and AHS are complex. The degree of AHS or IHS is not predictive of greater convergence when comparing a comment to a direct reply to that comment (i.e., when recovering the lexio-semantic content of parent comments). However, the strength of AHS and IHS on convergence is *strong* when comparing a comment that is high in either AHS or IHS to a sibling comment. Indeed, the effect of HS seems to be strongest when looking at threads of comments all replying to the same parent comment.

This result is surprising from the perspective of some theories of HS. Theories that posit that the primary illocutionary intent driving hate posting is to establish ingroup identity boundaries would predict that (as an example, see: Hoover et al., 2021), within self-selecting groups like the subreddits we studied, direct replies should elicit equal convergence as sibling comments. There should not have been any difference between parent and sibling comments expressing strong AHS or IHS. Both should be valid identity signals. While our comparison between the parameter estimates for the two is qualitative, that does not appear to be the case based on the current data. This does not invalidate social identity signaling explanations for the proliferation of HS, but it does complicate that picture.

At the same time, the observed information bottleneck arising in certain loci—this increase in the convergence between a hateful comment and sibling comments—makes sense if one views our results through the lens of "hate parties." In a hate party, like-minded members of groups post together in the same location in a larger discourse (Rea et al., 2024). While prior literature analyzing the hate party dynamic did not posit a potential signal for where partygoers should congregate, our results appear to indicate that the presence of a pre-existing, strongly hateful reply in some threads is a signal for others to post similar content within that same thread. *Older* sibling comments are a potential signal of where a party is taking place.

Regardless of whether one compares comments high in AHS or IHS to direct replies or to other sibling comments, the effect of AHS/IHS was strong (and significant) after October 7th, 2023, following the beginning of the most recent Israel–Palestine conflict. This change appears to indicate a general shift in how influential AHS and IHS are on subsequent comments. Prior research indicates that current affairs can affect group discourse, such that groups tend to coalesce on specific framings of those

events (Hiaeshutter-Rice and Hawkins, 2022). Our findings appear to validate this point. This is consistent with reporting by organizations tracking the rise of IHS and AHS online. But these results are perhaps more concerning. If AHS and IHS are *more influential* in online spaces since the start of the war in Gaza, this marks a stronger pull toward Antisemitic and Islamophobic sentiment when it is introduced online. Legislators, policymakers, and moderators should note that major, tragic events like those taking place in Gaza can strengthen the influence of AHS and IHS.

There are a few explanations for *why* AHS and IHS appear to predict higher convergence. One explanation could be that AHS and IHS are used *intertextually*. Gordon (2006) describes intertextuality as the capacity for language users to "creatively reshape language drawn from previous experiences in conversation" through interlocutors' use of repeated linguistic forms and motifs. As a device for collaborative semiosis, intertextuality emphasizes "how meaning is created in texts and interactions through callbacks to (and in anticipation of) other texts and interactions," (Gordon 2023) whose meaning has been normalized through shared group experiences. In any online community, whether that community is more tolerant of HS or not, "canned jokes, cartoons, and memes" can all come to be useful humorous, intertextual touchstones among members who share common ground and experiences (Tsakona and Chovanec, 2020). The use of humor as a mechanism for transmitting hateful content is well-attested in the literature on the history of HS in online spaces (Askanius, 2021; Ben-David and Fernández, 2016; Parvaresh, 2023; Rieger et al., 2021; Schmid, 2023). Prior qualitative research into the co-construction of hate-based content in Finnish and French has shown that intertextual artifacts at various levels of expression can serve as a mechanism for the covert expression of hateful content, thus avoiding potential moderation (Määttä, 2023). Widespread adoption of intertextual references could indeed decrease CE measurements across utterances if they are distributed widely enough within a given community.

Even if HS is couched in "humor," its use cannot be separated from group processes of meaning-making. In the same way that previous studies have shown that discourse around innocuous phenomena like referential speech acts concerning food are inextricably intertwined with "how posters use language to create meanings, negotiate morality and accountability, and construct identities" (Gordon, 2023) so too are hateful jokes, memes, and expressions a proposal for a referential Schelling point that other in-group users on social media can converge to.

Concerning the AHS and IHS, it could be that redditors are engaging in a kind of intertextuality—using snippets of prior topics, including prior AHS and IHS, to engage with the current state of discourse. If this is the case, it may help to explain why when the comment x expresses overt IHS or AHS it elicits greater convergence as well, as it explicitly ties the repetition of linguistic motifs to group affinity and shared history. That positive affinity then acts as a driver of increased convergence. Thus, intertextuality as an explanation would posit an additional reason or motivation for why such expressions end up having increased influence.

The current study was not equipped to test for the effects of intertextuality, but future studies may focus on such sociolinguistic elements to assess this hypothesis. Intertextual references are more difficult to identify when one is not a member of a given community, and our quantitative framework was not designed to identify it specifically. We the researchers are interlopers and observers within these particular subreddits, and we focused our study on convergence broadly, irrespective of how it is achieved. Thus, it is hard for us to identify intertextual references with high confidence. Even so, a few examples struck us as being potentially intertextual when reviewing the data. At one point, we observed that a member of one of the subreddits included in the AHS dataset wrote the following: "The jews fear the Samurai." While it was not entirely clear to us what was meant by the author, several subsequent posters commented with near identical comments, varying in minor iterations of capitalization. This incident, to us as outsiders, felt as though additional context was necessary to unpack it—it felt as though there was some history of prior discussions in the group that this phrase was tapping into, based on its lexical composition and the mimicry exhibited in the subsequent replies. Some piece of group lore that we were not privy to as outside observers. Indeed, upon further inspection of the group's history, we found reference to an entirely separate image board where posters had behaved identically with multiple commenters on the image board responding, one after another, with "The Jew fears the Samurai," and where the interpretation of the meme/screenshot as offered in a follow-up comment in the Reddit post indicated that the original post was meant to contrast the "honor" of Samurai culture with an antisemitic trope of a "dishonorable" Jew. While this was the most easily observed example we found of what appears to be ritualized language, surely others must be scattered throughout the data that are less salient and thus went unobserved. This example stood out because it bridged the gap from convergence broadly to linguistic mimicry specifically in a very salient way. While a single example is not conclusive evidence of widespread intertextuality within even a single community, it is an existential proof that points us in the direction of looking for it in subsequent work (likely using a very different methodological approach as well).

Regardless of the presence of intertextual references, the results we present here remain the same. Intertextuality, we posit, may be a possible contributor to our observations. It is an extremely interesting contributor with sociological implications. But just one of many possible rhetorical strategies yielding lower CE measurements between speakers.

Is it possible that the lower CE measurements we observe when comparing subsequent comments to prior, hateful comments could be due to members of these subreddits' irony or individuals mocking the progenitors of IHS and AHS by ironically instantiating prior redditors' hate-based ideas? If so, then CE would be powered not by true convergence, but by subsequent individuals conjuring pieces and ideas from prior comments to derogate them (Rossen-Knill and Henry, 1997; Sperber, 1984). While it is possible, that possibility was not borne out by the data.

If ironic mockery was a primary driver of lower CE values, then among the comments that had the lowest CE we would expect to see several comparisons wherein the comment y (1) refutes the comment x, and (2) does so while leveraging irony as a rhetorical device. To test this, we performed a targeted exploratory data analysis. We calculated the residual CE for comparisons of all comments x with a hatefulness rating ≥0.95 to subsequent comments y[4] and averaged the residual by dividing it by the number of tokens in the text x. We then sorted the data from lowest to highest average residual CE and manually assessed and counted the number of examples where the reply refuted the ideas put forth in the comment made by the progenitor of a hateful comment, for the first 100 rows of the data. We operationalized "refutation" as either (a) contradicting the main point made in the prior comment, or (b) consisting of an ad hominem attack against the original commenter themselves. We then assessed the comments that appeared to refute the original comment for elements of parody: (c) whether the second comment re-presents the ideas expressed in the first comment (which is captured by CE), (d) whether the second comment flaunts the text that it refutes,

and (e) whether the criticism appears to be couched in attempted comedy.[5] We repeated this process separately for AHS and IHS.

Within the AHS data, 9 out of the 100 examples met our criteria of refuting the point made in the comment *x*. Of those, only three appeared to have an ironic tone to us. A similar distribution of refutation existed within the IHS data. 10 out of the 100 replies analyzed appeared to refute the comment *x*. Of those, four appeared to use irony while doing so. Upon further study of the data, however, two of those four appeared to be using irony to *agree* with the original post, rather than refute it—there was a difference in the locutionary structure of these comments when compared to the possible illocutionary meaning of them. This left us with eight subsequent texts *y* that appear to refute the text *x*, of which two leveraged irony to do so.

Examples mentioned above from both the AHS and IHS data were too long to include in this paper or the supplementary materials. They have been included within the following OSF repository for others to view: https://bit.ly/AHS_IHS. None of the examples from either the AHS or IHS conditions were within the top 10 replies exhibiting the highest convergence.

Is it, thus, that lower CE measurements for more hateful comments are being driven by folks engaging in ironic mockery of one another in these subreddits? That does not seem to be the primary driver of CE measurements based on these exploratory results. Refutation is not common within replies that exhibit low convergence with the most overtly hateful comments in the data, and ironic refutation of them is even less so. In many ways, too, this is expected from the perspective of CAT. Convergence as a phenomenon is largely powered by prosocial affinity and ideological concordance. While ironic mockery leverages a faux veneer of convergence, it should be less common than individuals converging with one another earnestly in discourse.

It is worth mentioning as well that the majority of interactions occurring within the subreddits we studied would not meet the criterion for being offensive from the perspective of the subreddits' members. (1) If a group has a history of tolerating HS, then its deployment (especially when directed at out-group members), is unlikely to run afoul of group expectations or desires. (2) If HS is tolerated, then the linguistic formulae used to express it, including the specific kinds of HS used, the intensification strategy and diction of such utterances, and so forth, are unlikely to be associated with offensive speech. (3) All of these points are of course predicated on the degree to which the behavior is positively valenced within these communities (The points described here are adapted from the work of Culpeper, 2011).

The drop in CE associated with increasing measurements of AHS and IHS, and the evidence that the drop observed is not powered through refutation of the author's core points, both indicate that (3) is the most probable scenario/explanation for why CE is lower for some comments when those comments are higher in HS ratings. That also lends credence to points (1) and (2) above. While the current study is not equipped, nor was it our objective, to fully flesh out whether or not (1) and (2) are correct—unpacking whether those hold true requires intensive qualitative ethnographic work to uncover—the results practically call for such an extension of the current findings.

Does this mean that the behavior exhibited by members of these subreddits would not be offensive in a different context? No. But it is worth differentiating the behavior exhibited in an intragroup setting, like what is observed within these subreddits, versus behavior exhibited in an intergroup setting, like what we see on platforms like X, Bluesky, and others. The strongest examples of AHS and IHS we analyzed, based on the specific setting in which we observed them, can be understood as existing within a set of intragroup, social norms. As described in Walther (2024):

> "If hate proponents post messages in order to antagonize various targets, it stands to reason that the content of their messages should offer clearly hostile meanings to those targets and that they are posted in a way that is visible to those targets. However, at least two counterarguments exist. First, the anatomy of hate messages reveals considerable in-group symbology that is often unknown or uninterpretable by the targets. Second, hate mongers often retreat into more insulated virtual enclaves in which to post and share hate messages, where their supposed targets may be less and less likely to see their messages." (Walther, 2024)

Simply put, the examples of AHS and IHS we observe are for ingroup members, and ingroup members only. We are not observing "hate raids" or other forms of harassment, but the social behavior of individuals and their online peers.

It is worth noting that, in contrast to Rea et al. (2024) who studied hate parties in a more public "battlefield" of mainstream social media platforms, our study looks at the effects of AHS and IHS *within* communities of individuals who have been accused of being more tolerant to hate-based ideologies. It is from this vantage point that our results usefully expand on the dynamics of the "hate party" as a phenomenon. Our emphasis on haters interacting with members of their chosen online community on their home turf makes the finding that the hatefulness of the parent comment is not predictive of CE at all, all the more interesting. One could easily imagine a scenario in which the hatefulness of a comment signals to other Redditors that any additional hate-based comments posted in reply to the original comment a fair game. However, it is the presence of a hateful, prior *sibling* comment that appears to be the stronger signal of a good location for a hate party to commence in an intragroup setting. In some ways this is similar to the selection process of a location for a hate party when individuals engage with one another in *inter*group settings on other social media platforms—it is not that one member of a group necessarily produces the content that a hate party springs from, but rather that a set of individuals self-cluster into a sub-discourse at the same level of replies (i.e., in sibling comments) by replying to content or pre-existing comments on original content that is "congenial to hate comments" (Rea et al., 2024). Our results affirm that the signal for a hate party is not a social media user carving out a space online for others. A hate party gets off the ground when people post replies to the same content or comments expressing similarly hateful or hate-adjacent ideas. It is not a party until other like-minded people show up in the same comment thread.

A final, simple explanation may be that members of these groups *conventionalize* the ways that AHS and IHS are used in discourse. Lexical meaning is often dynamically co-negotiated by interlocutors during discourse (Christiansen and Chater, 2022; Clark and Wilkes-Gibbs, 1986; Enfield and Sidnell, 2022). This process of interaction and refinement of lexical meaning over time can yield highly conventionalized linguistic patterns, unique to the needs of interlocutors (Hawkins et al., 2019; Lupyan and Dale, 2016). It may well be that within the groups whose data we collected, repeated interactions have given way to conventionalized lexical patterns and meaning concerning the form of AHS and IHS. We leave this as a subject for future investigation.

Conclusion

We have shown through this study that as the degree to which comments strongly express either IHS or AHS increases, so too

does the influence of those comments on the subsequent posts of other Reddit users. Indeed, such speech acts have the perlocutionary effect of creating a bottleneck in the ways that members of several subreddits express their ideas in online interactions. This effect is strongest within particular loci—between comments that reply to the same parent comment.

There are a few limitations to the current study. We already mentioned that the word-vector model we used constrained our analyses to focus only on English text. In addition, the method deployed here can tell us about *how* information flows between comments in online interactions, but it is not equipped to tell us the content of those comments. Thus, while our results stand on their own as a warning about the influence of HS in online interactions, additional qualitative analysis is required to study the exact patterns that are passed from comment to comment. Despite this general inference, it may nevertheless represent a concerning outcome. One can easily imagine redditors or other individuals espousing hateful content, and using that as a strategy to simply increase the influence of their ideas overall. A final limitation stems from the way we identified HS, IHS, and AHS. Trends in how HS is expressed online change rapidly and often via *covert* rather than overt expression (Bhat and Klein, 2020). This makes the identification of HS difficult even with up-to-date lexical resources like the Weaponized Word's lexicons, or by machine learning-based methods like the HS classifier we used. This makes it easy to miss a subset of examples of HS in a study like ours.

Despite these limitations, our results indicate that examples of overt AHS and IHS exert a significant force on individuals' diversity of expression in the communities studied.

Our findings may have implications for content moderation or monitoring where it is practiced by organizations in online spaces. It has been common moderation practice to delete only those comments online that have been flagged as overtly hateful or marked as problematic by other users while ignoring subsequent comments entirely. Our results indicate, however, that the influence of any single comment espousing strong Islamophobic or Antisemitic sentiment is much more diffuse. Indeed, concerning at least IHS and AHS, *if* organizations are concerned with the creep of ideas that might cause individuals to take up violence against Jewish and Muslim communities, and how that might affect their brand identity, special attention should be paid to other comments in the immediate context of the flagged comment as well, to look for traces of similarly hateful content. Consistent with the observations of YouTube comments made by Rea et al. (2024), information flow is highest between comments espousing AHS/IHS and other, sibling comments in their immediate vicinity. Inwood and Zappavigna (2023) found that when studying the proliferation of hateful comments on YouTube, individuals grouped around specific "personae" of individuals who acted as loci for espousing particular kinds of calls to action—some commenters flocked to "educators" who dispelled conspiracy myths, while others commented alongside "inciters" who actively called for violence against targeted groups, for example. We did not test for this in our current study—we did not posit a mechanism for doing such. But this, too, could motivate why we observed low entropy among sibling comments, especially when the older sibling comment contained IHS/AHS. The emerging perspective based on both the prevailing literature and our current results, thus, indicates that quarantining single messages does not effectively curtail the influence of AHS or IHS.

As a final thought, while HS is a harmful attack on the psychological well-being of its explicit targets, we would like to make the case that HS is equally an attack on the agency of those engaged in discourse before its introduction. When a community member introduces hateful content to conversation, they rob everyone else of some of the expressiveness they would have otherwise been able to exercise, as seen in the sharp decrease in convergence-entropy we observe. If everyone else gets sucked into a "hate party", then those users who wanted to express a greater diversity of ideas are left alienated within the communities that they hoped would be more welcoming to them as individuals. We hope that pointing this out may empower community members whose discourse and ideas have been derailed by the introduction of HS to metaphorically reclaim the floor that has otherwise been taken from them.

Data availability

Reddit's terms of service for the use of their API prohibits the sharing of raw text data but provides for sharing indexes of data analyzed, including submission identifiers (IDs). Thus, we can provide submission IDs for all the submissions analyzed in the study with valid proof of CITI certification to "rehydrate" the raw text data. To support replication efforts, the code used to collect data is provided at https://bit.ly/AHS_IHS. To directly replicate the LME analyses in this study, an anonymized and shuffled CSV file containing the factors studied and CE values is available at https://bit.ly/AHS_IHS.

Code availability

All of our code for both data collection and analysis is available on GitHub for individuals to use at their leisure. The GitHub repository for the project is also linked to our OSF repo here: https://bit.ly/AHS_IHS.

Received: 13 August 2024; Accepted: 21 February 2025;
Published online: 02 April 2025

Notes

1 The examples provided in this paper were taken from the data collected as described in our methods and materials. However, several ethical issues arise when publishing direct quotations from users' social media profiles (Norman Adams, 2024). The texts presented in this paper have been altered using an LLM to retain the core ideas expressed in each example to protect the original user's right to privacy. Because LLMs are prone to generating incorrect or even unrelated text (Xu et al., 2024)—i.e., "hallucinations"—all measurements reported, including CE values in examples ex: refute - ex: hs-not-convergent were calculated using the original, unaltered texts.

2 Lexicographic data courtesy of the Weaponized Word (weaponizedword.org). Weaponized Word curates an up-to-date lexical database of HS and extremist vocabulary.

3 This meant that there was a gradual sloping decrease in probability for values, such that ~62% of the total cumulative probability density function fell within a range of [0, 1.5] (CoE values are restricted to a range from [0, 2]).

4 To calculate the residual, we left out the terms in our regression equation related directly to HS. These left variables for the number of tokens in the comment x, n_x, the number of tokens in the comment y, n_y, the number of comments interceding the comment x and the comment y, *commentΔ*, the user who wrote the comment x, the user who wrote the comment y, and the specific subreddit that the comments were posted to. We then subtracted the predicted value from the regression using these terms from the observed value to calculate the residual. To control for how much the residual could be distributed across multiple tokens in an utterance, we divided the residual by n_x to calculate the average residual for an utterance.

5 Combinatively, these factors are consistent with the model of parody expressed in Rossen-Knill and Henry (1997). We did not assess comments based on whether or not they were *comedic*. Comedy is a perlocutionary effect, and thus not fully in the control of an utterance's author. However, we did look for indicators of attempted comedic effect, like violations of Grice's Maxims of relation and quantity that could render an utterance inference vulnerable. Comments that change the topic unexpectedly but do not refute the claims made in the comment x or rebuke/ridicule the author of x do not count as direct refutation.

References

ADL (2023) U.S. antisemitic incidents skyrocketed 360% in aftermath of attack in Israel, according to latest ADL Data Anti-Defamation League (2024) New

York, https://www.adl.org/resources/press-release/us-antisemitic-incidents-skyrocketed-360-aftermath-attack-israel-according
Alfonseca K (2023) Los Angeles area plagued by antisemitic attacks in 'tsunami' of hate: advocates. ABC News. https://www.ajc.org/reports-and-emblematic-examples-of-antisemitic-hate-speech-and-violence-since-october-7
Allison I (2024) CAIR: New data shows the end of 2023 was a 'Relentless' wave of bias, community resilience is 'impressive'. https://www.cair.com/press_releases/cair-new-data-shows-the-end-of-2023-was-a-relentless-wave-of-bias-community-resilience-is-impressive/
Alviar C, Kello CT, Dale R (2023) Multimodal coordination and pragmatic modes in conversation Lang Sci 97:101524. https://doi.org/10.1016/j.langsci.2022.101524
Askanius T (2021) On frogs, monkeys, and execution memes: exploring the humor-hate nexus at the intersection of Neo-Nazi and alt-right movements in Sweden Telev New Media 22(2):147–165. https://doi.org/10.1177/1527476420982234
Austin J (1975) How to do things with words. 2nd edn. Harvard Univ Press, Cambridge
Bäck EA, Bäck H, Sendén MG, Sikström S (2018) From I to we: group formation and linguistic adaption in an online xenophobic forum. J Soc Polit Psychol 6(1):76–91
Beknazar-Yuzbashev G, Jiménez Durán R, McCrosky J, Stalinski M (2022) Toxic content and user engagement on social media: evidence from a field experiment. https://doi.org/10.2139/ssrn.4307346
Ben-David A, Fernández AM (2016) Hate speech and covert discrimination on social media: monitoring the Facebook pages of extreme-right political parties in Spain Int J Commun 10:1167–1193
Bhat P, Klein O (2020) Covert hate speech: White nationalists and dog whistle communication on Twitter. In G Bouvier, JE Rosenbaum (ed) Twitter, the public sphere, and the chaos of online deliberation, Springer Int Publishing, Cham, pp 151–172. https://doi.org/10.1007/978-3-030-41421-4
Bradac JJ, Mulac A, House A (1988) Lexical diversity and magnitude of convergent versus divergent style shifting: perceptual and evaluative consequences. Lang Commun 8(3):213–228
Brennan SE, Clark HH (1996) Conceptual pacts and lexical choice in conversation. J Exp Psychol Learn Mem Cogn 22(6):1482–1493
Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. (2020) Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
Calvert C (1997) Hate speech and its harms: a communication theory perspective. J Commun 47(1):4–19
Christiansen MH, Chater N (2022) The unbearable lightness of meaning. The language game: how improvisation created language and changed the world. Basic Books, New York
Cinelli M, Pelicon A, Mozetič I, Quattrociocchi W, Novak PK, Zollo F (2021) Dynamics of online hate and misinformation Sci Rep 11(1):22083
Clark HH, Wilkes-Gibbs D (1986) Referring as a collaborative process. Cognition 22(1):1–39
Coco MI, Dale R, Keller F (2018) Performance in a collaborative search task: the role of feedback and alignment. Top Cogn Sci 10(1):55–79
Collen D (2023) "Let Them Kill Each Other": the Israel-Palestine war is the perfect storm for Canada's far-right. *GNET*. https://gnet-research.org/2023/10/30/let-themkill-each-other-the-srael-palestine-war-is-the-perfect-storm-forcanadas-far-right/
Culpeper J (2011) Impoliteness: using language to cause offence. Number 28 Stud Interact Sociolinguist Cambridge Univ Press, Cambridge
Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein J, Doran C, Solorio T (Eds), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
Enfield NJ, Sidnell J (2022) Intersubjectivity, activity, accountability. in consequences of language: from primary to enhanced intersubjectivity. The MIT Press, Cambridge
Gamble J (2024) CAIR records 'staggering' 3578 complaints of bias in the US since October 7. CNN
Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B (2022) Shared computational principles for language processing in humans and deep language models. Nature Neurosci 25(3):369–380
Gordon C (2006) Reshaping prior text, reshaping identities. Text Talk - Interdisci J Lang, Discourse Commun Stud26(4/5):545–571
Gordon C (2023) *Intertextuality 2.0: metadiscourse and meaning-making in an online community*. Oxford Univ Press, New York, 1st ed. https://doi.org/10.1093/oso/9780197642689.001.0001
Hawkins RX, Goodman ND, Goldstone RL (2019) The emergence of social norms and conventions. Trends Cogn Sci 23(2):158–169
Henson B, Reyns BW, Fisher BS (2013) Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. J Contemp Crim Justice 29(4):475–497
Hiaeshutter-Rice D, Hawkins I (2022) The language of extremism on social media: an examination of posts, comments, and themes on Reddit. Front Polit Sci 4:805008
Hickey D, Fessler DMT, Lerman K, Burghardt K (2024) The peripatetic hater: predicting movement among hate subreddits. arXiv:2405.17410 [cs]
Hilte L (2023) How is linguistic accommodation perceived in instant messaging? A survey on teenagers' evaluations and perceptions. J Lang Soc Psychol 42(4):431–452
Hoover J, Atari M, Mostafazadeh Davani A, Kennedy B, Portillo-Wightman G, Yeh L (2021) Investigating the role of group-based morality in extreme behavioral expressions of prejudice. Nat Commun 12(1):4585
Inwood O, Zappavigna M (2023) Conspiracy theories and white supremacy on YouTube: exploring affiliation and legitimation strategies in YouTube comments. Social Media + Soc 9(1):20563051221150410
Jones S, Cotterill R, Dewdney N, Muir K, Joinson A (2014) Finding Zelig in text: a measure for normalising linguistic accommodation. In: Proceedings of the 25th international conference on computational linguistics: technical papers (COLING 2014), Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp 455–465
Leets L, Giles H (1999) Harmful speech in intergroup encounters: an organizational framework for communication research. Ann Int Commun Assoc 22(1):91–137
Lewandowski N, Jilka M (2019) Phonetic convergence, language talent, personality and attention. Front Commun 4 https://doi.org/10.3389/fcomm.2019.00018
Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov, V (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692
Lupyan G, Dale R (2016) Why are there different languages? The role of adaptation in linguistic diversity Trends Cogn Sci 20(9):649–660. https://doi.org/10.1016/j.tics.2016.07.005
Määttä SK (2023) Linguistic and discursive properties of hate speech and speech facilitating the expression of hatred: evidence from Finnish and French online discussion boards Internet Pragm 6(2):156–172. https://doi.org/10.1075/ip.00094.maa
Manson JH, Bryant GA, Gervais MM, Kline MA (2013) Convergence of speech rate in conversation predicts cooperation. Evol Human Behav 34(6):419–426
Marshal MP, Dietz LJ, Friedman MS, Stall R, Smith HA, McGinley J(2011) Suicidality and depression disparities between sexual minority and heterosexual youth: a meta-analytic review J Adolescent Health 49(2):115–123. https://doi.org/10.1016/j.jadohealth.2011.02.005
Muir K, Joinson A, Cotterill R, Dewdney N (2016) Characterizing the linguistic chameleon: personal and social correlates of linguistic style accommodation Hum Commun Res 42(3):462–484. https://doi.org/10.1111/hcre.12083
Nishida S, Blanc A, Maeda N, Kado M, Nishimoto S (2021) Behavioral correlates of cortical semantic representations modeled by word vectors PLOS Comput Biol 17(6):e1009138. https://doi.org/10.1371/journal.pcbi.1009138
Norman Adams N (2024) 'Scraping' Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of Covid-19 Int J Soc Res Methodol 27(1):47–62. https://doi.org/10.1080/13645579.2022.2111816
Owen T (2023) Neo-Nazis and the far-right are trying to hijack pro-Palestine protests. https://www.vice.com/en/article/k7zx5a/neo-nazis-hijack-pro-palestine-protest-mike-enoch
Ozalp S, Williams ML, Burnap P, Liu H, Mostafa M (2020) Antisemitism on Twitter: collective efficacy and the role of community organisations in challenging online hate speech Social Media + Soc 6(2):2056305120916850. https://doi.org/10.1177/2056305120916850
Parvaresh V (2023) Covertly communicated hate speech: a corpus-assisted pragmatic study J Pragmat 205:63–77. https://doi.org/10.1016/j.pragma.2022.12.009
Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue Behav Brain Sci 27(2):169–190. https://doi.org/10.1017/S0140525X04000056
Pluta A, Mazurek J, Wojciechowski J, Wolak T, Soral W, Bilewicz M (2023) Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain Sci Rep 13(1):4127. https://doi.org/10.1038/s41598-023-31146-1
Project TT (2023) TTP White supremacists on X premium use Israel-Hamas conflict to push hate agenda. https://www.techtransparencyproject.org/articles/white-supremacists-on-x-premium-use-israel-hamas-conflict
Rae L (2012) Beyond belief: pragmatics in hate speech and pornography. In I Maitra, MK McGowan, editors, Speech and harm: controversies over free speech, Oxford Univ Press. https://doi.org/10.1093/acprof:oso/9780199236282.003.0004
Rea S, Mathew B, Kraemer J (2024) "Hate parties": networked antisemitism from the fringes to YouTube. In JB Walther, RE Rice, editors, Social processes of

online hate, 1st ed, Routledge, London, pp 168–192. https://doi.org/10.4324/9781003472148
Reports and Emblematic Examples of Antisemitic Hate Speech and Violence Since October 7 American Jewish Committee (2023) New York
Rieger D, Kümpel AS, Wich M, Kiening T, Groh G (2021) Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media*, p 14, https://royalsocietypublishing.org/doi/10.1098/rspb.2014.0488
Rosen ZP, Dale R (2023) BERTs of a feather: Studying inter- and intra-group communication via information theory and language models. Behav Res Methods 56(4):3140–3160 https://doi.org/10.3758/s13428-023-02267-2
Rossen-Knill DF, Henry R (1997) The pragmatics of verbal parody. J Pragmat 27(6):719–752
Saha K, Chandrasekharan E, De Choudhury M (2019) Prevalence and psychological effects of hateful speech in online college communities. Proc ACM Web Sci Conf 2019:255–264
Schmid UK (2023) Humorous hate speech on social media: a mixed-methods investigation of users' perceptions and processing of hateful memes. New Media Soc, 14614448231198169. https://doi.org/10.1177/14614448231198169
Searle JR (1975) A taxonomy of illocutionary acts. Language, mind, and knowledge, vol 7. University of Minnesota Press, Minneapolis, pp 344–369
Shannon CE, Weaver W (1949) The mathematical theory of communication. Univ Illinois Press. Google-Books-ID: IZ77BwAAQBAJ. https://www.degruyter.com/document/doi/10.1515/ijsl.1984.46.49/html
Shatz I (2017) Fast, free, and targeted: Reddit as a source for recruiting participants online. Social Sci Comp Rev 35(4):537–549
Soliz J, Giles H, Gasiorek J (2021) Communication accommodation theory: converging toward an understanding of communication adaptation in interpersonal relationships. In DO Braithwaite, P Schrodt, editors, Engaging theories in interpersonal communication: multiple perspectives, 3rd ed Routledge, New York, pp 130–142. https://doi.org/10.4324/9781003195511
Sperber D (1984) Verbal irony: pretense or echoic mention? J Exp Psychol Gen 113(1):130–36. https://doi.org/10.1080/13645579.2022.2111816
Srivastava S, Wentzel SD, Catala A, Theune M (2025) Measuring and implementing lexical alignment: A systematic literature review. Comput Speech Lang 90(1):101731. https://doi.org/10.1016/j.csl.2024.101731
Tsakona V, Chovanec J (2020) Revisiting intertextuality and humour: fresh perspectives on a classic topic Eur J Humour Res 8(3):1–15. https://doi.org/10.7592/EJHR2020.8.3.Tsakona
Tynes BM, Giang MT, Williams DR, Thompson GN (2008) Online racial discrimination and psychological adjustment among adolescents J Adolescent Health 43(6):565–569. https://doi.org/10.1016/j.jadohealth.2008.08.021
Vidgen B, Thrush T, Waseem Z, Kiela D (2021) Learning from the worst: dynamically generated datasets to improve online hate detection. arXiv:2012.15761 [cs]
Walther JB (2022) Social media and online hate Curr Opinion Psychol 45:101298. https://doi.org/10.1016/j.copsyc.2021.12.010
Walther JB (2024) Making a case for a social processes approach to online hate. In Social processes of online hate. Routledge. 28
Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A et al. (2020). Transformers: State-of-the-Art Natural Language Processing. In Liu Q., Schlangen D. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6
Xu Z, Jain S, Kankanhalli M (2024) Hallucination is inevitable: an innate limitation of large language models. arXiv:2401.11817 [cs]

Author contributions

ZPR designed the research. RD guided research questions surrounding the interpersonal dynamics in groups studied and provided critical feedback throughout. ZPR and RD jointly collaborated on the statistical procedures and analyses reported on in this study.

Competing interests

The authors declare no competing interests.

Ethical approval

All data used in this study is publicly available and exempt from requiring IRB approval.

Informed consent

The current study used publicly available texts posted to social media and did not require informed consent.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-04647-9.

Correspondence and requests for materials should be addressed to Z. P. Rosen or Rick Dale.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025